Process Data Applications in Educational Assessment

Jitong Qi

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy under the Executive Committee of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023
Abstract

Process Data Applications in Educational Assessment

Jitong Qi

The widespread adoption of computer-based testing has opened up new possibilities for collecting process data, providing valuable insights into the problem-solving processes that examinees engage in when answering test items. In contrast to final response data, process data offers a more diverse and comprehensive view of test takers, including construct-irrelevant characteristics. However, leveraging the potential of process data poses several challenges, including dealing with serial categorical responses, navigating nonstandard formats, and handling the inherent variability. Despite these challenges, the incorporation of process data in educational assessments holds immense promise as it enriches our understanding of students’ cognitive processes and provides additional insights into their interactive behaviors. This thesis focuses on the application of process data in educational assessments across three key aspects.

Chapter 2 explores the accurate assessment of a student’s ability by incorporating process data into the assessment. Through a combination of theoretical analysis, simulations, and empirical study, we demonstrate that appropriately integrating process data significantly enhances assessment precision.

Building upon this foundation, Chapter 3 takes a step further by addressing not only the target attribute of interest but also the nuisance attributes present in the process data to mitigate the issue of differential item functioning. We present a novel framework that leverages process data as
proxies for nuisance attributes in item response functions, effectively reducing or potentially eliminating differential item functioning. We validate the proposed framework using both simulated data and real data from the PIAAC PSTRE items.

Furthermore, this thesis extends beyond the analysis of existing tests and explores enhanced strategies for item administration. Specifically, in Chapter 4, we investigate the potential of incorporating process data in computerized adaptive testing. Our adaptive item selection algorithm leverages information about individual differences in both measured proficiency and other meaningful traits that can influence item informativeness. A new framework for process-based adaptive testing, encompassing real-time proficiency scoring and item selection is presented and evaluated through a comprehensive simulation study to demonstrate the efficacy.
Table of Contents

Acknowledgments .................................................. vii

Chapter 1: Introduction and Background ......................... 1
  1.1 Item Response Theory ........................................ 1
    1.1.1 Dichotomous Response Models ......................... 1
    1.1.2 Polytomous Response Models ......................... 3
    1.1.3 Latent Trait Estimation ................................. 5
  1.2 PIAAC PSTRE Data ........................................ 6
  1.3 Process Feature Extraction ................................ 9
    1.3.1 Multidimensional Scaling Feature Extraction ........... 11
    1.3.2 Autoencoder Feature Extraction ....................... 13
  1.4 Thesis Overview ........................................... 15

Chapter 2: Accurate Assessment via Process Data .................. 16
  2.1 Introduction ............................................... 16
  2.2 Latent Trait Estimation with Processes and Responses ........ 18
    2.2.1 Statistical Formulation ............................... 18
    2.2.2 Procedure ........................................... 20
    2.2.3 Theoretical Analysis ................................. 21
Chapter 4: Computerized Adaptive Testing via Process Data

4.1 Introduction

4.2 Computerized Adaptive Testing Algorithm via Proficiency Responses

4.2.1 Heteroscedastic Estimate Model

4.2.2 Real-time Proficiency Estimation

4.2.3 Model Calibration

4.2.4 Adaptive Item Selection

4.3 Simulation Studies

4.3.1 Experiment Settings

4.3.2 Comparison and Evaluations

4.3.3 Results

4.4 Conclusion

References

Appendix A: Proofs of Theorem 1 and Theorem 2

Appendix B: Proof of Theorem 3
## List of Figures

1.1 Home page of the PSTRE sample item. ................................. 7

1.2 Web page returned from clicking the second link (i.e., “Work links”) on the home page. ................................. 8

1.3 Web page returned from clicking “Learn More” on the “Work links” website. . . . 9

1.4 Structure of autoencoder .................................................. 13

1.5 Structure of RNN-based autoencoder .................................. 14

2.1 Response-based (baseline) and process-incorporated estimators’ agreement with true $\theta$ across different sample sizes. ............................................... 26

2.2 Box plots of MSE in different strata of true $\theta$ ($N = 2000$). ................................. 28

2.3 Simulation results of experiment II. ................................. 29

2.4 Response-based (baseline) and process-incorporated estimators’ MSE with respect to reference-set ability estimate on the PIAAC data. $x$—axis is the number of items used for computing $\hat{\theta}_X$ and $\hat{\theta}_Y$ in the scoring set. Box plots represent the distribution of MSE across the 50 partitions of $B_s$ and $B_r$. ................................. 31

2.5 Distribution of response-based (baseline) and process-incorporated Kendall’s $\tau$ with reference set performance across 50 partitions on the PIAAC data. ................................. 32

2.6 Distribution of MSEs with reference set latent trait estimate in each residual decile. Box plots are the distributions of MSEs within each decile bin across the 50 partitions of scoring and reference sets. ................................. 34

2.7 Scatterplot of process-incorporated and response-based $\theta$ estimates with 14 items. . 36

3.1 Example model structure without intercepts. ................................. 48
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>RMSE and correlation between true ability and estimation with or without DIF reduction under uniform DIF</td>
</tr>
<tr>
<td>3.3</td>
<td>RMSE and correlation between true ability and estimation with or without DIF reduction under non-uniform DIF</td>
</tr>
<tr>
<td>3.4</td>
<td>Distribution of examinee self-reported age for the PSTRE assessment data</td>
</tr>
<tr>
<td>3.5</td>
<td>Absolute correlation between first principal feature and logarithm of sequence length</td>
</tr>
<tr>
<td>3.6</td>
<td>Observed $L^2$ distance and the hypothesis testing thresholds as a function of feature dimension</td>
</tr>
<tr>
<td>3.7</td>
<td>Maximum, 0.9 quantile and median of response function distances in log scale</td>
</tr>
<tr>
<td>3.8</td>
<td>Fisher information of parameter $a_j$ in log scale</td>
</tr>
<tr>
<td>3.9</td>
<td>Sample mean of Fisher information of $\theta$ in log scale</td>
</tr>
<tr>
<td>3.10</td>
<td>Empirical distributions and means (dashed vertical lines) of nuisance ability estimates, separated by email-moving strategies</td>
</tr>
<tr>
<td>3.11</td>
<td>Empirical distribution and mean of nuisance ability estimation for both groups</td>
</tr>
<tr>
<td>3.12</td>
<td>Scatterplot of the IRT-based latent trait estimates ($x$-axis) and the DIF-corrected latent trait estimates ($y$-axis)</td>
</tr>
<tr>
<td>4.1</td>
<td>Comparison of CAT algorithms in Scenario 1</td>
</tr>
<tr>
<td>4.2</td>
<td>Comparison of CAT algorithms in Scenario 2</td>
</tr>
<tr>
<td>4.3</td>
<td>Comparison of CAT algorithms in Scenario 3</td>
</tr>
<tr>
<td>4.4</td>
<td>Comparison of average test overlap rates</td>
</tr>
<tr>
<td>4.5</td>
<td>Comparison of exposure rate</td>
</tr>
</tbody>
</table>
List of Tables

1.1 Descriptive information of the 14 PIAAC PSTRE items. . . . . . . . . . . . . . . . . 10

3.1 Descriptive statistics of the 13 PIAAC problem-solving items. . . . . . . . . . . . 58

3.2 Out-of-sample accuracy for binary response prediction using MDS features. . . . . 60

3.3 DIF detection results using logistic regression and $L_2$ distance method. . . . . . . 62
Acknowledgements

First and foremost, I would like to express my gratitude to my advisors, Prof. Zhiliang Ying and Prof. Jingchen Liu. Their guidance throughout this research journey have been invaluable. I also extend my thanks to my collaborators: Susu Zhang, Xueying Tang, Zhi Wang and Ling Chen. Together, we made the accomplishments in this thesis possible.

I owe a special debt of gratitude to my parents. Their unwavering support and understanding have always been my backbone. Their love and encouragement have been my safe harbor, and their faith in me has fueled my every endeavor.

I also wish to thank my hamster, Dazhan. During the demanding days of my Ph.D., he consistently brought moments of peace and joy to my life, always being the adorable companion that lightened my mood.

Last but certainly not least, a heartfelt thanks to my girlfriend, Yuanlu Bai. From our first meeting to becoming the closest of friends, from understanding each other to falling deeply in love, we took the bold step of embarking on an entirely new journey together. Throughout these five years of our Ph.D. journey, we’ve faced challenges and celebrated joys side by side. This period in New York will forever be cherished and remembered as some of the most precious years of my life. As we move forward, I eagerly anticipate the new adventures and memories we’ll craft together.
Chapter 1: Introduction and Background

1.1 Item Response Theory

Item response theory (IRT) has become a cornerstone of modern psychological measurement as it provides a framework for understanding how both individual differences in trait levels and the properties of test items influence a person’s item responses [1]. Various IRT models are developed for different types of responses, which are often dichotomous (correct/incorrect) or polytomous (partial score). In this section, we present an overview of classical IRT models. Specifically, we denote the final response of examinee $i$ to item $j$ as $Y_{ij}$, and the underlying target latent trait as $\theta_i$.

1.1.1 Dichotomous Response Models

This section provides an overview of various IRT models that are suitable for analyzing binary data, ranging from simple to complex models. It should be noted that in some cases, polytomous responses can be transformed into binary responses without significant loss of information [2, 3]. However, if the binary scoring results in a loss of information about the underlying latent trait, a polytomous model in Section 1.1.2 should be used instead.

One-Parameter Logistic Model or Rasch Model

The model assumes that the probability of an examinee $i$ giving a correct answer to item $j$ is described by the following equation:

$$P(Y_{ij} = 1|\theta_i) = \frac{e^{\theta_i - b_j}}{1 + e^{\theta_i - b_j}}$$  \hspace{1cm} (1.1)
where $b_j$ is the difficulty parameter of item $j$. The Rasch model [4] is also known as the one-parameter logistic (1PL) model because it only contains one parameter for each item. The logit transformation of the probability in Equation (1.1) gives the difference between the examinee’s latent trait ability and the item difficulty parameter.

**Two-Parameter Logistic Model**

The two-parameter logistic (2PL) model is an extension of the Rasch model that includes item discrimination parameters. In this case, the probability of a person solving item $j$ is given by:

$$P(Y_{ij} = 1|\theta_i) = \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}}$$

Here, $a_j$ is the item discrimination parameter for item $j$, which captures how well the item differentiates between individuals with different levels of the latent trait. The 2PL model is particularly useful for measurements where items are not equally related to the latent trait, or where some items are more informative than others in terms of a person’s standing on the latent trait. The inclusion of discrimination parameters makes the 2PL model more flexible and allows for a more accurate estimation of individuals’ trait levels.

**Three-Parameter Logistic Model**

The three-parameter logistic (3PL) model takes the 2PL model one step further by adding a parameter to represent an item characteristics curve that does not fall to zero. For example, in multiple-choice cognitive items, where guessing is possible, the probability of success is substantially greater than zero, even for low trait levels. The 3PL model accommodates guessing by adding a lower-asymptote parameter $c_j$, as shown below:

$$P(Y_{ij} = 1|\theta_i) = c_j + (1 - c_j) \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}}.$$  \hspace{1cm} (1.3)

It is important to note that if the lower asymptotes is independent for each item, it may lead to
non-identifiable issues [1]. Therefore, a common lower asymptote is often estimated for all items or for groups of similar items.

**Normal Ogive Models**

The normal ogive model uses a different function than the logistic function to produce the item characteristic curve, which is the cumulative proportion of cases in the normal distribution. Specifically, the normal ogive model gives the item response function as:

\[
P(Y_{ij} = 1 | \theta_i) = \int_{-\infty}^{w_{ij}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt
\]

(1.4)

where \( w_{ij} \) can be \( \theta_i - b_j \) or \( a_j(\theta_i - b_j) \) depending on the number of parameters. Analogous to the 3PL model, a lower asymptote may be added to the normal ogive model to account for guessing in item response data. [5] provides a detailed comparison of the logistic models and ogive models.

### 1.1.2 Polytomous Response Models

Numerous psychological measurement include items with multiple ordered responses as they are more informative and reliable than dichotomously scored items. These types of items provide more variability in the response patterns and can better distinguish individuals with different levels of the latent trait. This section presents several popular polytomous response models. For a more comprehensive literature review, interested readers are referred to [6]. Throughout the rest of this section, we use \( M_j \) to denote the maximal score of item \( j \) and thus, the number of response categories is \( M_j + 1 \) for item \( j \).

**Graded Response Model**

The graded response model (GRM, [7]) is appropriate to use when item responses can be characterized as ordered categorical responses. The GRM is a generalization of the 2PL model described in Section 1.1.1 and falls under the ”difference mode” rubric [8]. To be specific, the
GRM assumes that the probability of obtaining a response higher or equal than \( m \) is

\[
P(Y_{ij} \geq m | \theta_i) = \frac{e^{a_j(\theta_i - b_{jm})}}{1 + e^{a_j(\theta_i - b_{jm})}}
\]  

(1.5)

where \( m = 1, \ldots, M_j \). Then the item response function is given by

\[
P(Y_{ij} = 0 | \theta_i) = 1 - P(Y_{ij} \geq 1 | \theta_i) \quad (1.6)
\]

\[
P(Y_{ij} = m | \theta_i) = P(Y_{ij} \geq m | \theta_i) - P(Y_{ij} \geq m + 1 | \theta_i) \quad \text{for} \quad m = 1, \ldots, M_j - 1
\]  

(1.7)

\[
P(Y_{ij} = M_j | \theta_i) = P(Y_{ij} \geq M_j | \theta_i).
\]  

(1.8)

**Partial Credit Model**

The partial credit model (PCM, [9]) is particularly well-suited for analyzing test items that require multiple steps in the solution process, such as math problems, and when partially correct answers are possible. The item response function of the PCM is given by

\[
P(Y_{ij} = y_{ij} | \theta_i) = \frac{\exp \sum_{m=0}^{y_{ij}} (\theta_i - b_{jm})}{\sum_{r=0}^{M_j} \exp \sum_{m=0}^{r} (\theta_i - b_{jm})}.
\]  

(1.9)

One way to interpret the PCM model is to consider the conditional probability of obtaining the higher category score given a choice between that category and the preceding one, e.g.,

\[
P(Y_{ij} = m | Y_{ij} = m \text{ or } Y_{ij} = m - 1) = \frac{e^{\theta_i - b_{jm}}}{1 + e^{\theta_i - b_{jm}}}.
\]  

(1.10)

The right-hand side of Equation (1.10) has the same structure as the item response function of the Rasch model described in Section 1.1.1. Therefore, the PCM can be considered as an extension of the Rasch model. Furthermore, \( b_{jm} \) represents the item step difficulty associated with a category score of \( m \). The higher the value of a particular \( b_{jm} \), the more difficult a particular step is relative
to other steps within an item.

The PCM can be further extended to allow for different slopes of the item response functions. This generalized version is known as the generalized partial credit model (GPCM, [10]). To be specific,

$$P(Y_{ij} = y_{ij} | \theta_i) = \frac{\exp \sum_{m=0}^{y_{ij}} a_j (\theta_i - b_{jm})}{\sum_{r=0}^{M_j} \exp \sum_{m=0}^{r} a_j (\theta_i - b_{jm})}. \quad (1.11)$$

1.1.3 Latent Trait Estimation

This section provides an overview of how IRT models are utilized to derive examinee scores on a latent trait variable. It should be noted that the following discussion is not restricted to scoring dichotomous or polytomous IRT models. Therefore, we denote the IRT model parameters of item $j$ as $\zeta_j$.

The maximum likelihood estimator (MLE) is defined as

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \sum_{j} \log(P(Y_j = y_j | \theta, \hat{\zeta}_j)), \quad (1.12)$$

where $\hat{\zeta}_j$ represents the calibrated item parameters.

In addition, two other common estimators under the Bayesian framework are the Bayesian expected a posteriori estimator (EAP) and the Bayesian modal estimator (BME), given by:

$$\hat{\theta}_{\text{EAP}} = E[\theta | Y] \quad \text{and} \quad \hat{\theta}_{\text{BME}} = \arg \max_{\theta} P(\theta | Y), \quad (1.13)$$

where $P(\theta | Y) \propto p(\theta) \prod_j P(Y_j = y_j | \theta, \hat{\zeta}_j)$ with $p(\theta)$ being the prior distribution [e.g., 11].
1.2 PIAAC PSTRE Data

Carried out by the Organization for Economic Co-operation and Development (OECD), the Programme for International Assessment of Adult Competency (PIAAC) [e.g., 12] is an international survey of the cognitive and workplace skills of working-age individuals around the world. The first cycle of the PIAAC survey in 2012 assessed three cognitive skills, namely literacy, numeracy, and problem-solving in technology rich environments (PSTRE), on participants from 24 countries and regions with ages between 16 and 65 years. In addition to the three cognitive assessments, the participants were further surveyed on their demographic background and other information related to their occupation and education.

The current study focuses on the PSTRE assessment, where individuals were administered a series of computer-based interactive items. PSTRE ability refers to the ability to use digital technology, communication tools, and the internet to obtain and evaluate information, communicate with others, and perform practical tasks [13]. Successful completion of the PSTRE tasks requires both problem-solving skills and familiarity with digital environments. The test environment of each item resembled commonly seen informational and communicative technology (ICT) platforms, such as e-mail clients, web browsers, and spreadsheets. Test takers were asked to complete specific tasks in these interactive environments. Individuals’ entire log of interactions with each item was recorded as log data. In addition, based on the extent of task completion, polytomous final scores were derived for each item.

A sample item that resembles PSTRE tasks is shown in Figure 1.1. Respondents can read the task instructions on the left side and work on the task in the simulated interactive environment on the right. This item requires respondents to identify, from the five web pages presented on the screen, all pages that do not require registration or fees and to bookmark them. By clicking on each link, they will be redirected to the corresponding website where they can learn more. For example, clicking “Work Links” directs them to Figure 1.2, and further clicking on “Learn More” directs them to the page on Figure 1.3. Once having finished working on the task, a test taker can click on
the right arrow (“Next”) on the bottom-left. A pop-up window will ask them to confirm their decision by clicking “OK” or to return to the question by clicking “Cancel”. A respondent who clicked on the aforementioned two links, bookmarked the page using the toolbar icon, and moved on to the next question will have the recorded action sequence of “Start, Click_W2, Click_Learn_More, Toolbar_Bookmark, Next, Next_OK”.

Figure 1.1: Home page of the PSTRE sample item.
Reprinted from OECD Sample Questions and Questionnaire.

The computer-based version of the 2012 PIAAC survey randomly assigned each respondent with two blocks of cognitive items, where each block consisted of a fixed set of items that assessed either literacy, numeracy, or PSTRE proficiency. This thesis uses the PSTRE response and process data of individuals from five countries and regions, including the United Kingdom (England and
Northern Ireland, Ireland, Japan, the Netherlands, and the United States of America, who were assigned to PSTRE for both blocks. The five countries and regions were relatively similar in performance distribution on the PIAAC PSTRE assessment. Each PSTRE block consisted of 7 items, and thus the two blocks total to 14 items. Note that a recorded action sequence of “Start, Next, Next_OK” indicates that the test taker did not perform any action on the item and moved on to the next question. This type of behavior can be regarded as omission and is distinguished from either credited or uncredited responses. For each item, the action sequence of each test taker was recorded, and a polytomous final score calculated based on predefined scoring rubrics was available. PIAAC uses the final scores together with other demographic covariates to obtain plausible values of individuals’ proficiency on PSTRE. Table 1.1 presents descriptive information.
of the 14 PSTRE items, including the task names and the descriptive statistics of the final scores and action sequences. For each item, around 26,000 responses were recorded while to report the summary, we excluded individuals who omitted any of the 14 items, resulting in a total of 2304 test takers who responded to all 14 PSTRE items.

1.3 Process Feature Extraction

Recent advancements in computer-based testing have made it possible to collect process data, which document the problem-solving processes that examinees go through to arrive at their final responses. In contrast to final response data, process data are more heterogeneous and contain valuable insights into problem-solving strategy, nuisance ability, etc. Recent research has demonstrated
Table 1.1: Descriptive information of the 14 PIAAC PSTRE items.

<table>
<thead>
<tr>
<th>Item ID</th>
<th>Task name</th>
<th>Final Score</th>
<th>Sequence Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Score levels</td>
<td>Median</td>
</tr>
<tr>
<td>U01a</td>
<td>Party Invitations</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>U01b</td>
<td>Party Invitations</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>U02</td>
<td>Meeting Room</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>U03a</td>
<td>CD Tally</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>U04a</td>
<td>Class Attendance</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>U06a</td>
<td>Sprained Ankle</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>U06b</td>
<td>Sprained Ankle</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>U07</td>
<td>Book Order</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>U11b</td>
<td>Locate Email</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>U16</td>
<td>Reply All</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>U19a</td>
<td>Club Membership</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>U19b</td>
<td>Club Membership</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>U21</td>
<td>Tickets</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>U23</td>
<td>Lamp Return</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

Note. Descriptive statistics calculated based on the 2304 participants without omission; Score levels: number of ordinal response categories; Action types: the number of possible actions in the log data; Sequence length: the number of actions recorded in a test taker’s process data.

The usefulness of process data in various applications, including classifying problem-solving strategies, interpreting different action patterns, and assessing the proficiency in the trait of interest [14, 15, 16].

However, process data are often in a nonstandard format that is difficult to analyze directly. Thus, it is a reasonable approach to preprocess the data by embedding each response process to a finite-dimensional vector space to address this challenge. There are multiple methods to fulfill this task, including \( n \)-gram language modeling [14], sequence-to-sequence autoencoder [17], and multidimensional scaling [18]. In [19], actions that are not essential for solving the question were recoded into aggregate-level categories and further used to predict the final response based on early-window clickstreams. [20] proposed a method based on the longest common sequence to summarize the process information and identify behavioral patterns. [21] further generalized it to incorporate timestamp information to construct edge-weighted undirected graph. A more detailed literature review can be found in [22]. In this section, we focus on the multidimensional scaling and autoencoder methods.
In this thesis, let $S_i = (S_{i1}, \ldots, S_{iL})$ denote the sequence of actions, where $S_{il}$ is the $l$-th action and $L$ is the total number of steps.

### 1.3.1 Multidimensional Scaling Feature Extraction

Multidimensional scaling (MDS) is a popular dimension reduction technique and data visualization tool utilized across many disciplines. MDS seeks to embed objects in a vector space based on their pairwise dissimilarities, such that similar objects are located close to one another and dissimilar objects are positioned far apart. Consequently, the effectiveness of MDS is heavily influenced by the choice of distance metric used in the analysis, as it should be appropriate for the specific dataset under consideration.

As the response data consists of discrete categorical sequences, arithmetic calculations are not applicable. Moreover, response processes from different respondents can vary widely in length, so standard distance measures may not be appropriate. Instead, an appropriate dissimilarity measure should capture the variation among response processes. Furthermore, the order of actions is crucial and should be taken into consideration, as it can reflect the problem-solving strategies employed by the respondents and provide other useful information.

Based on the aforementioned considerations, we adopt a dissimilarity measure called ordering-based sequence similarity (OSS, [23]). Consider $S_m$ and $S_n$ be two action sequences to be compared. The dissimilarity between $S_m$ and $S_n$ is defined as follows:

$$
\frac{L_m + L_n}{f(S_m, S_n) + g(S_m, S_n)},
$$

where $f(S_m, S_n)$ quantifies the dissimilarity among the actions that appear in both $S_m$ and $S_n$, and $g(S_m, S_n)$ is the count of actions appearing in only one of $S_m$ and $S_n$. The sequence lengths are $L_m$ and $L_n$, respectively.

To clarify the technical details of the OSS dissimilarity measure, we provide a precise definition. Let $S^a$ denote the sequence consisting of chronologically ordered positions of action $a$ in sequence $S$, and let $L^a$ be the length of $S^a$, i.e., the number of times that $a$ appears in $S$. We use
$S^a(k)$ to denote the position of the $k$-th appearance of $a$ in $S$. For two sequences $S_m$ and $S_n$, let $C_{mn}$ denote the set of actions that appear in both $S_m$ and $S_n$, and $U_{mn}$ denote the set of actions that appear in $S_m$ but not in $S_n$, and vice versa. Then $f(S_m, S_n)$ and $g(S_m, S_n)$ are defined as

$$f(S_m, S_n) = \frac{\sum_{a \in C_{mn}} \sum_{k=1}^{K_{mn}^a} |S_m^a(k) - S_n^a(k)|}{\max(L_m, L_n)}$$

and

$$g(S_m, S_n) = \sum_{a \in U_{mn}} L_m^a + \sum_{a \in U_{nm}} L_n^a$$

where $K_{mn}^a = \min(L_m^a, L_n^a)$. To summarize, the adopted dissimilarity measure takes into account the variation among response processes, the unequal sequence lengths across respondents, and the order of actions that can reflect problem-solving strategies and other relevant information.

After computing the symmetric dissimilarity matrix $D = (d_{mn})$ using the definition in Equation (1.14), Equation (1.15), and Equation (1.16), MDS map each action sequence to a latent vector $x$ in the $K$-dimensional Euclidean space $\mathbb{R}^K$, with the aim of minimizing the objective function

$$\sum_{m<n} (d_{mn} - ||x_m - x_n||)^2$$

with respect to $x_m \in \mathbb{R}^K$, which represents the latent vector of the action sequence $S_m$. Several algorithms have been proposed to solve this optimization problem, including stochastic gradient descent and the BFGS method.

Note that the latent feature dimension $K$ is a crucial hyper-parameter, and it can be selected through $n$-fold cross-validation. Additionally, to obtain interpretable features, principal component analysis (PCA) can be applied to the resulting MDS embedding.
1.3.2 Autoencoder Feature Extraction

Autoencoder is a type of artificial neural network that has been widely used in various fields, such as computer vision, speech recognition, and natural language processing. Its primary objective is to learn a compressed and meaningful representation of the input data. In other words, it reduces the dimensionality of the input data by projecting it into a lower-dimensional space, while retaining its essential features. The key idea behind autoencoder is to train the network to reproduce the input data as accurately as possible, by compressing it into a compact representation and then reconstructing it back to its original form.

![Figure 1.4: Structure of autoencoder](image)

As illustrated in Figure 1.4, autoencoder consists of two parts, the encoder and the decoder, which are trained together using backpropagation algorithm by stochastic gradient descent (SGD). The encoder transforms the input data into a compressed representation, while the decoder reconstructs the original data from the compressed representation. Since the low-dimensional vector contains sufficient information to restore the original data and is in a standard and simpler format, autoencoders are commonly used for dimensionality reduction and feature extraction.

The encoder and decoder in an autoencoder are typically implemented as neural networks, such as recurrent neural networks (RNNs), long-short-term-memory (LSTMs), or gate recurrent unit (GRUs). In this section, we focus on RNNs as an example to illustrate a candidate model. The structure of the RNN autoencoder is depicted in Figure 1.5. The encoder has three steps. First, each unique action is embedded as a numeric vector in $\mathbb{R}^K$, so an input action sequence is transformed into a sequence of $K$-dimensional embeddings. Second, an RNN is used to sequentially summarize
the information in the embedding sequence up to a time step. Lastly, the last vector in the output of the encoder RNN is retained as the feature vector. The decoder of the RNN autoencoder also has three steps. First, the feature vector is repeated to form a sequence of vectors. Second, the sequence of vectors is passed into another RNN to obtain another sequence of vectors. Each vector in this latter sequence contains the information of the action at the corresponding time step. In the last step, \( S_l \) is predicted for \( l = 1, \ldots, L \) using a multinomial logit model. The loss function used to measure the discrepancy between an action sequence \( S \) and its reconstructed version \( \hat{S} \) is the average cross entropy, i.e.,

\[
l(S, \hat{S}) = -\frac{1}{L} \sum_{l=1}^{L} \sum_{a} S_{l}^{a} \log(\hat{S}_{l}^{a}). \tag{1.18}
\]

In practice, neural networks are often over-parametrized, which increases the risk of overfitting the data. Therefore, it is common to stop the algorithm before convergence to prevent overfitting. One popular technique for early stopping is to monitor the value of the objective function on a validation set. The algorithm is run for a sufficient number of iterations, and the parameter values producing the lowest monitored objective function value are selected as the final estimates. This approach helps to avoid overfitting the model and to achieve better generalization performance.
1.4 Thesis Overview

This thesis explores three important research topics in educational assessments using process data. The first topic, presented in Chapter 2, focuses on accurately assessing a student’s ability by including process data in the assessment. While traditional assessments rely on final responses, process data collected by computer-based interactive items provide additional information about a student’s detailed interactive processes. Through theoretical, simulated, and empirical data, Chapter 2 demonstrates that appropriately including such information in the assessment can significantly enhance relevant assessment precision.

Chapter 3 considers reducing differential item functioning via process data. Intuitively, differential item functioning arises when item responses depend not only on the target attribute of interest but also a number of unobserved nuisance attributes whose distributions are heterogeneous among subpopulations. Since process data contain rich information concerning each individual, including cognitive characteristics, social background, etc., it provides new possibility to measure the nuisance attributes. Chapter 3 presents a novel framework to reduce or possibly completely remove differential item functioning by including process data as proxies of nuisance attributes in the item response functions. The proposed framework is implemented and evaluated on both simulated data and real data from the PIAAC PSTRE items.

Finally, in Chapter 4, the potential for using process data in computerized adaptive testing is explored, where the adaptive item selection leverages information about individual differences both on the measured proficiency and on other meaningful traits that can influence item informativeness. This chapter presents a new framework for computerized adaptive testing that incorporates process data information in both real-time proficiency scoring and item selection. The efficacy of the process-based adaptive testing algorithm is evaluated via a simulation study.
Chapter 2: Accurate Assessment via Process Data

2.1 Introduction

The main task of educational assessment is to provide reliable and valid estimates of test takers’ ability based on their responses to test items. Much of the efforts in the past decades focused on the IRT models. These models have been designed to provide various approaches to model different types of responses as described in Section 1.1.

The rapid advancement of information technology has enabled the collection of various sorts of process data from assessments, ranging from reaction times on multiple-choice questions to the log of problem-solving behavior on computer-based constructed-response items. In particular, the sequence of actions performed by test takers to solve a task, which documents the problem-solving process, can contain valuable information on top of final responses, that is, dichotomous or polytomous scores on how well the task was completed.

Process data often contain rich information about the test takers. Much of the literature focuses on developing new research directions. In this chapter, we try to answer the question of how existing research could benefit from the analysis. Specifically, we develop a method to incorporate the information in process data into the scoring formula. There are two key features to consider in measurement, reliability and validity. The current chapter focuses on the improvement in reliability: We show that the process-incorporated latent ability estimate improves measurement accuracy. In particular, we demonstrate through simulation and empirical analyses that the process-incorporated scoring rule yields much higher reliability than the IRT-model-based scoring rule which is outlined in Section 1.1.3. On the current empirical data, score based on two items’ process data on average could be as accurate as that based on final scores on five items. Furthermore, we provide a theoretical framework under which process-incorporated scores yield lower
measurement error of test takers’ abilities under certain regularity conditions.

One particular challenge in process-incorporated scoring is that process data record the entire problem-solving process, are highly unstructured and variable, and reveal different aspects of a test taker, including construct-irrelevant characteristics. It is unclear which part of the process is related to the particular trait of interest. Conventionally, it is up to the domain experts to identify construct-relevant process features and derive scoring rules.

The proposed approach, on the other hand, considers an automated method: Features are extracted through an exploratory analysis that typically lacks interpretation. We take advantage of the IRT score, which serves as a guide for the scoring rule to yield an estimator for the measured trait. The entire procedure does not require particular knowledge of the item design.

As shown in the PIAAC PSTRE example, process data are frequently in a nonstandard format, which can make it challenging to analyze them directly. To overcome this issue, we preprocess the data by embedding each response process into a finite-dimensional vector space. Multiple methods have been proposed to achieve this goal, as described in Section 1.3. In this chapter, we adopt multidimensional scaling to accomplish this task. By utilizing this approach, we aim to provide a more comprehensive understanding of the process data and obtain more accurate results from our subsequent analyses.

In the psychometrics literature, plenty of research has been conducted on the use of additional information, such as response times, to improve measurement accuracy [e.g., 24]. To the authors’ best knowledge, this is the first piece of work to use problem-solving log data to improve measurement accuracy. A literature that is remotely related to the current work is automated scoring systems for constructed responses. Extensive research has been conducted on automated scoring of essays, which aims at producing essay scores comparable to human scores based on examinees’ written text [e.g., 25, 26, 27, 28]. Other than essay scoring, automated scoring engines have been developed for scenario-based questions in medical licensure exams [29], constructed-response mathematics problems [30], speaking proficiency exams [31], and prescreening for post-traumatic stress disorder based on self-narratives [32, 33].
Many of these systems were shown to produce comparable scores to expert ratings. Readers are referred to [34] and [35] for comprehensive reviews of the history, applications, conceptual foundations, and validity considerations of automated scoring systems.

The proposed approach differs from most automated scoring systems in its objective. Whereas automated scoring systems are often designed to reproduce expert- or rubric-derived scores in an automated and standardized manner, the purpose of the proposed Rao-Blackwellization approach is not to reproduce the final scores but to refine the latent trait estimates based on original final scores with the additional information from the problem-solving process.

The rest of the chapter is organized as follows. Section 2.2 describes the statistical formulation. The proposed method for process-incorporated score refinement is introduced. Theoretical results on mean squared error (MSE) reduction in latent trait estimation are presented. Section 2.3 reports the results of simulation studies that verify the theoretical findings. Section 2.4 presents an empirical example on the PSTRE assessment in the 2012 PIAAC survey, where the proposed method is compared to original response-based scoring in several aspects. A discussion of the implications and limitations is provided in Section 2.5.

### 2.2 Latent Trait Estimation with Processes and Responses

This section presents a generic framework for process-incorporated latent trait measurement. Section 2.2.1 describes a statistical formulation that the current framework is built upon. Section 2.2.2 introduces the proposed Rao-Blackwellization approach and the specific procedures. Section 2.2.3 provides a theoretical analysis of the process-incorporated trait estimator.

#### 2.2.1 Statistical Formulation

Consider a test of $J$ items designed to measure a latent trait, $\theta$. For an examinee, on each item $j$, both the final item response and the action sequence for problem-solving are recorded. Denote the item response by $Y_j$, which can be a polytomous score ranging between 0 and $C_j$, representing different degrees of task completion. Further denote the action sequence by $S_j = (S_{j1}, \ldots, S_{jL_j})$,
where $L_j$ is the total number of actions performed on the item, and $S_{jl}$ is the $l$th action.

We consider the case where the action sequences record problem-solving details and no randomness is involved in the grading. Therefore, the process response contain at least as much information as the final outcomes since the final item response can be derived from the action sequence through a deterministic scoring rule $f$ such that $Y_j = f(S_j)$. Further suppose that the final responses to the $J$ items are conditionally independent given $\theta$ and follow some item response function in Section 1.1,

$$P(Y_j = y_j | \theta, \zeta_j),$$

where $\zeta_j$ is the parameters of item $j$.

For the present purpose of latent trait estimation, we assume that the item parameters ($\zeta_j$s) have been calibrated and only the latent trait $\theta$ is unknown. Denote the pre-calibrated parameters of item $j$ by $\hat{\zeta}_j$. The latent trait $\theta$ for each individual can be estimated based on the final response from one or more items. Details of commonly used latent trait estimators are discussed in Section 1.1.3.

We aim at refining the $\theta$ estimator with a procedure that makes use of process data. Since action sequences are in a non-standard format, instead of working directly with $S_j$, we work with the $K-$dimensional numerical features extracted from $S_j$, denoted $X_j = (X_{j1}, \ldots, X_{jk}) \in \mathbb{R}^K$. There are no restrictions on the feature extraction method except that the produced features $X_j$ must preserve the full information on the final response $Y_j$, in other words, $\sigma(Y_j) \subseteq \sigma(X_j)$, where $\sigma(\cdot)$ denotes the $\sigma-$algebra generated by the random variable. Intuitively, this requires the extracted features to preserve full information about the final score, such that the latter can be perfectly predicted by the process data features. The feature extraction methods presented in Section 1.3 can be readily applied depending on practical settings. Even if the process does not contain complete final outcome information, the final outcome can always be added as an additional dimension of the process features, to guarantee $\sigma(Y_j) \subseteq \sigma(X_j)$. 

10
2.2.2 Procedure

Let \( X = (X_1, \ldots, X_J) \) denote the process features from all \( J \) items and \( X_{-j} \) the process features from the \( J - 1 \) items excluding item \( j \). Denote the latent trait estimate based on all \( J \) final responses by \( \hat{\theta}_Y \), which can be obtained through the estimators in Section 1.1.3. Further, let \( \hat{\theta}_{Y_j} \) be the estimator derived from a single response outcome \( Y_j \), for instance, using the EAP estimator in Equation (1.13).

The final response-based trait estimator, \( \hat{\theta}_Y \), can be refined by the following procedures that incorporate process features. The new estimator is denoted by \( \hat{\theta}_X \).

**Procedure 1** (Construction of process-incorporated estimator).

1. For each \( j = 1, \ldots, J \):
   - Regress \( \hat{\theta}_{Y_j} \) on \( X_{-j} \) to obtain \( T_{X_{-j}} = E[\hat{\theta}_{Y_j} | X_{-j}] \).
   - Regress \( \hat{\theta}_Y \) on \( T_{X_{-j}} \) and \( Y_j \) to obtain \( \hat{\theta}_{X_{-j}} = E[\hat{\theta}_Y | T_{X_{-j}}, Y_j] \).

2. Compute the overall process-incorporated estimator, \( \hat{\theta}_X = \frac{1}{J} \sum_{j=1}^{J} \hat{\theta}_{X_{-j}} \).

In practice, the explicit distributions of \( \hat{\theta}_{Y_j} | X_{-j} \) and \( \hat{\theta}_Y | T_{X_{-j}}, Y_j \) are unknown. The two conditional expectations, \( E[\hat{\theta}_{Y_j} | X_{-j}] \) and \( E[\hat{\theta}_Y | T_{X_{-j}}, Y_j] \) in Procedure 1 are approximated on finite samples using generalized linear models. Alternatively, deep neural networks may be considered to capture nonlinear relationships. Although \( J \) regressions are required for both steps, the implementation may be paralleled to make it computationally efficient.

For illustration, consider a test of \( J \) binary items administered to \( N \) respondents. For respondent \( i \) and item \( j \), let \( s_{ij} \) and \( Y_{ij} \in \{0, 1\} \) denote the response process and the response outcome, respectively. Suppose that the response outcomes follow a two parameter logistic model in Section 1.1.1.

The following steps provide a roadmap to implement Procedure 1.

1. IRT parameter estimation: Fit the 2PL model on the binary outcomes \( \{Y_{ij} : i = 1, \ldots, N, j = \)
1, . . . , J} to obtain the item parameter estimates \( \{ \hat{\zeta}_j = (\hat{a}_j, \hat{b}_j) : j = 1, \ldots, J \} \) using marginal maximum likelihood estimation.

2. Process feature extraction: For each item \( j \), extract features \( X_{1j}, \ldots, X_{Nj} \) from the problem-solving processes \( S_{1j}, \ldots, S_{Nj} \). The MDS method in Section 1.3.1 or the action sequence autoencoder in Section 1.3.2 implemented in the ProcData R package [36] can be used for this step.

3. Response-based (baseline) latent trait estimation: One can choose from the commonly used estimators described in Section 1.1.3. For each respondent \( i \) and item \( j \), get the single-item trait estimate \( \hat{\theta}_{i,Y_j} \) based on \( Y_{ij} \) and \( \hat{\zeta}_j \). Additionally, based on examinee \( i \)’s final responses to all \( J \) items, \( Y_i = (Y_{i1}, \ldots, Y_{iJ}) \), and \( \hat{\zeta} = (\hat{\zeta}_1, \ldots, \hat{\zeta}_J) \), obtain \( \hat{\theta}_{i,Y} \).

4. First conditional expectations: For each \( j \), fit a regression \( \hat{\theta}_{Y_j} \sim X_{-j} \) based on \( \{(X_{i(-j)}, \hat{\theta}_{i,Y_j}) : i = 1, \ldots, N\} \) to approximate \( E[\hat{\theta}_{Y_j}|X_{-j}] \) and calculate the fitted values \( \{T_{i,X_{-j}} : i = 1, \ldots, N\} \). For example, one can use ridge regression [37] with shrinkage parameter selected by cross-validation.

5. Second conditional expectations: For each \( j \), fit another regression \( \hat{\theta}_X \sim (T_{X_{-j}}, Y_j) \). One simple choice is ordinary least squares with \( (1, T_{X_{-j}}, Y_j, T_{X_{-j}}Y_j) \) as predictors, where \( T_{X_{-j}}Y_j \) is the interaction term.

6. Averaging step: The average of the fitted values \( \{\hat{\theta}_{i,X_{-j}} : j = 1, \ldots, J\} \) in step (5) is the final process-incorporated trait estimate, \( \hat{\theta}_{i,X} \), for respondent \( i \).

2.2.3 Theoretical Analysis

In this subsection, we provide theoretical justifications for this procedure. In particular, we elaborate a set of assumptions under which the process data score improves latent trait estimation in terms of mean squared error.
The first assumption requires the conditional expectation of $\hat{\theta}_{Yj}$ given $\theta$ to be monotonically increasing. This assumption is satisfied by well-designed cognitive items and latent trait estimators.

A1. (Monotonicity assumption) $m_j(\theta) = E[\hat{\theta}_{Yj}|\theta]$ is monotone in $\theta$ and has a finite second moment.

Secondly, we assume that the response outcome of item $j$ is correlated with an individual’s behaviors on other items only through the measured trait $\theta$ and not through other latent or observed traits. Since the process features can include rich information other than the measured trait, this local independence assumption requires $Y_j$ to be “good”, in the sense that no construct-irrelevant, persistent traits affect final performance. In other words, measurement error comes from sources of random, instead of systematic error [38]. For example, the process features $X_{-j}$ may reflect a respondent’s computer usage habits, such as whether they tend to double- or single-click on buttons. However, the final score, $Y_j$, shall not differentiate individuals with different clicking habits, as long as they have the same level of $\theta$. We do allow $Y_j$ to be very “rough” measurements, in other words, the measurement error can be large, as long as it is due to random instead of systematic variations. Similarly, $\hat{\theta}_{Yj}$ can be biased and can have a large standard error, as long as the monotonicity assumption (A1) is satisfied.

A2. (Local independence assumption) Conditioning on the latent trait $\theta$, $Y_j$ and $X_{-j}$ are independent.

Finally, we consider the distribution of the process features, $X_{-j}$, given the measured trait. Note that the problem-solving process can depend on traits other than $\theta$. These unobserved, construct-irrelevant traits are assumed random and integrated out from the probability density, thus resulting in the conditional density of $X_{-j}$ given $\theta$. We impose the usual exponential family assumption on process features for technical development. This is equivalent to assuming the existence of a unidimensional sufficient statistic of $\theta$ independent of sample size [39]. The natural parameter $\eta_j(\theta)$ is assumed to be monotone so that there is no identifiability issue for $\theta$. 
A3. (Exponential family assumption) The probability density function for features $X_{-j}$ for each $j$ takes the following form,

$$f(X_{-j}|\theta) = \exp\{\eta_j(\theta) T_j(X_{-j}) - A_j(\theta)\} h_j(X_{-j}), \quad (2.1)$$

where $T_j(X_{-j})$ is a sufficient statistic for $\theta$ and the natural parameter $\eta_j(\theta)$ is monotone in $\theta$ with a finite second moment.

Theorem 1 shows that the first step of our proposed procedure summarizes the process data features to sufficient statistics.

**Theorem 1.** **Under Assumptions A1–A3, $T_{X_{-j}}$ is a sufficient statistic of $X_{-j}$ for $\theta$.**

Based on the sufficiency of $T_{X_{-j}}$, we can further show that $\hat{\theta}_X$ reduces the MSE of $\hat{\theta}_Y$, as stated in Theorem 2. The proof of this result uses the Rao-Blackwell theorem [40, 41] and also shows that every $\hat{\theta}_{X_{-j}}$ produced by step 2 of the procedure removes conditional variance and improves $\hat{\theta}_Y$ in terms of MSE. The proofs of Theorem 1 and Theorem 2 are provided in the Appendix.

**Theorem 2.** *If assumptions A1-A3 hold for all $J$ items, then*

$$E[(\hat{\theta}_X - \theta)^2|\theta] \leq E[(\hat{\theta}_Y - \theta)^2|\theta] \quad \text{for every } \theta. \quad (2.2)$$

The MSE reduction by incorporating process features implies an increase in statistical efficiency for estimating $\theta$. Putting Theorem 2 in the psychometric context, the MSE reduction of $\theta$ estimator translates to the reduction of conditional standard error of measurement at all possible proficiency levels, therefore, higher measurement reliability. In practice, this allows one to derive more reliable scores (i.e., latent trait estimates) for individuals in assessments. Alternatively, by using the procedure to incorporate process data, one is able to achieve comparable measurement precision to traditional outcome-based scoring with fewer items.

Here, we explain the rationale behind the assumptions and expected consequences of assumption violation: Assumption A1 is the most important among all. It provides a crucial guide to
extracting the relevant part of the process information on the measured trait. This also suggests
that the final-response-based IRT model yields a valid estimator with random noise, written as
\[ \hat{\theta}_{Y_j} = m_j(\theta) + \varepsilon_j. \] Assumption A2 ensures that \( \varepsilon_j \) is not predictable by the process data of other
items, and furthermore \( \varepsilon_j \) has zero mean across the population. Thus, if A2 is not valid, we
may expect a certain amount of bias introduced to the process estimator. A3 provides a technical
framework for us to discuss efficiency. We choose the natural exponential family because it is the
first-order approximation of a large class of parametric families.

2.3 Simulation Studies

Simulation studies were conducted to compare outcome- and process-incorporated estimators
of the latent trait and we focus on the leave-one-out estimator since theoretical guarantees are
provided for it.

2.3.1 Experiment Settings

We generated respondents’ latent trait \( \theta_1, \ldots, \theta_N \) independently from the standard normal dis-
tribution. Given examinee \( i \)'s underlying true \( \theta_i \), the response process and the final response to
each item were then generated, both dependent on the latent trait \( \theta_i \). Specifically, for respondent \( i \)
and item \( j \), the response outcome \( Y_{ij} \) followed a Rasch model in Section 1.1.1.

To generate the problem-solving process, we considered a Markov model and an action set of
26 English letters. Each letter represents 1 of 26 possible actions recorded at each time point. The
probability transition matrix was distinct for each respondent-item pair and denoted as \( P^{(ij)} =
(p^{(ij)}_{kl})_{1 \leq k,l \leq M} \) for the \( i \)th respondent and the \( j \)th item. Given the probability transition matrix,
we generated an action sequence starting from “A”, where the subsequent actions were sampled
according to \( P^{(ij)} \) until the final state “Z” appears. Excluding the column for “A” and the row for
“Z”, the upper right \((M - 1) \times (M - 1)\) submatrix of \( P^{(ij)} \) was computed according to

\[
p^{(ij)}_{kl} = \frac{\exp(\theta_i u^{(j)}_{kl})}{\sum_{r=1}^{M-1} \exp(\theta_i u^{(j)}_{kr})},
\]

(2.3)
where \( (u_{kl}^{(j)})_{1 \leq k, l \leq M - 1} \) were generated independently from Uniform\((-10, 10)\) for each item. Each item’s action sequences hence depended on the underlying \( \theta \) through the transition probabilities between actions. Note that, by the current simulation design, the final response, generated separately, was not a function of the process and hence not perfectly predictable from process. To ensure the perfect predictability of response from process features on item \( j \), we included the final response as an additional dimension for the process features in subsequent analysis.

Two experiments were devised to evaluate the effect of sample size \( N \) and test length \( J \). Experiment I considers four different sample sizes: \( N = 200, 500, 1000 \) and \( 2000 \). The number of items \( J \) was fixed to three with difficulty parameters \( b_1 = 0, b_2 = 1, b_3 = -1 \) in Equation (1.1). Each condition of \( N \) was replicated 100 times.

Experiment II considers different test lengths. We considered a maximum of 20 items and generated the difficulty parameter \( b_j \) from Uniform\((-1, 1)\) for each item. Starting from two items, we added one more observed item for estimation in each step until all 20 items were included. The sample size \( N \) was fixed at 2000.

MDS features were extracted from the response process on each item, with latent dimension \( K \) chosen by five-fold cross-validation from candidate values \( 10, 20, \ldots, 50 \). To guarantee perfect predictability of \( Y_j \) by \( X_j \), the final response to each item was added as an additional dimension to the process features. Following the illustration in Section 2.2.2, we first estimated the item parameters \( b_1, b_2, b_3 \) by marginal maximum likelihood estimation. Then, we used response outcomes to calculate the baseline EAP estimator, \( \hat{\theta}_Y \), as well as the single-item response-based EAP estimators, \( \hat{\theta}_{Y_1}, \ldots, \hat{\theta}_{Y_J} \). Note that by using the EAP, we minimize the posterior MSE.

Ridge regression [42] was used for the first conditional expectation, \( E[\hat{\theta}_{Y_j} | X_{-j}] \), and the shrinkage parameter was tuned to minimize the deviance in five-fold cross-validation. For the second conditional expectation, \( E[\hat{\theta}_Y | T_{X_{-j}}, Y_j] \), we regressed \( \hat{\theta}_Y \) on \( (1, T_{X_{-j}}, Y_j, T_{X_{-j}}Y_j) \) by ordinary least squares.
2.3.2 Results

The estimators $\hat{\theta}_Y$ and $\hat{\theta}_X$ were evaluated by two criteria, MSE and Kendall’s rank correlation $\tau$ [43]. The MSE of an estimator $\hat{\theta}$ was calculated by

\[
\text{MSE}(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} (\hat{\theta}_i - \theta_i)^2,
\]

(2.4)

where $\hat{\theta}_i$ is the estimate for the $i$-th respondent, and $\theta_i$ is the true latent trait of respondent $i$. The Kendall’s $\tau$ between estimated and true $\theta$ can also be calculated for both estimators. In contrast to MSE, Kendall’s $\tau$ considers the extent to which the estimated ranking aligns with the true ranking of latent trait, which is the interest of norm-referenced tests.

![Figure 2.1: Response-based (baseline) and process-incorporated estimators’ agreement with true $\theta$ across different sample sizes.](image)

As shown in Figure 2.1, the process-incorporated estimator ($x$-axis) outperformed the outcome-based estimator ($y$-axis) in terms of both MSE and Kendall’s $\tau$. In each subplot, the 100 points correspond to results from 100 replications. For smaller sample sizes of $N = 200$ or 500, the MSE
of the process-incorporated estimator was higher than that of the response-based estimator in some replications. As the sample size increased, the proposed procedure consistently achieved smaller MSE.

The Kendall’s $\tau$ of the process-incorporated estimator was consistently higher than that produced from the baseline estimator across sample sizes and replications, and the improvement became more substantial when $N$ increased. As shown in the subplot for $N = 2000$, with three items, $\tau$ could increase from around 0.45 to over 0.60 after applying the procedure to incorporate process information.

Figure 2.2 takes a closer look at MSE by dividing respondents into 10 groups based on their true latent trait, with $N = 2000$. Group 1 has the lowest $\theta$ and group 10 has the highest $\theta$. Within each group, we calculated MSE for baseline and process-incorporated estimators over the 100 simulated data sets. The box plots show that for groups 2 to 9, the process-incorporated estimator substantially reduced MSE. For examinees with extreme true $\theta$s (groups 1 and 10), both estimators had larger errors.

Figure 2.3 displays the results of experiment II where test length was considered. In the left panel, the MSE of the process-incorporated estimator (red line) remained below that of the outcome-based estimator (black line) as the number of items increased from 2 to 20. The proposed procedure reduced the MSE by over a half. The improvement in Kendall’s $\tau$ was also consistent across different test lengths as shown in the right panel. For instance, when $J = 7$, Kendall’s $\tau$ rose from 0.5 to 0.7 after the refinement.

### 2.4 Empirical Example: PIAAC PSTRE

The proposed approach for score refinement was further applied to the data collected from the PSTRE assessment from the 2012 PIAAC survey. The empirical analyses were guided by two overarching objectives. First, the performance of the response- and process-incorporated latent trait estimators were compared, similar to the simulation studies. Second, because response-based and process-incorporated estimators are expected to produce different latent ability estimates for
the same examinee, we further examined the problem-solving patterns associated with large discrepancies in response- and process-incorporated $\hat{\theta}$s. In the following subsections, a description of the PIAAC PSTRE data considered in this chapter is first provided, followed by the methods and findings from the empirical analyses.

2.4.1 Overall Performance in Latent Proficiency Estimation

**Evaluation Methods**

With empirical data, respondents’ true $\theta$s were unknown. The two classes of proficiency estimators were instead compared on their agreement with performance on a separate set of items designed to measure the same trait. Specifically, the 14 PSTRE items were split into two sets of 7 items. One set of 7 items, denoted the scoring set ($B_s$), was used to obtain the response- and
the process-incorporated estimators $\hat{\theta}^{(s)}$ and $\hat{\theta}^{(s)}$. A separate latent trait estimate, $\hat{\theta}^{(r)}$, can be obtained from the final responses to the remaining 7 items, denoted the reference set ($B_r$). Any trait estimate obtained from the scoring set does not use reference set response information, and $\hat{\theta}^{(r)}$ serves as an external criterion for evaluating $\hat{\theta}^{(s)}$ and $\hat{\theta}^{(s)}$.

The 14 items can be partitioned into scoring and reference sets in $\binom{14}{7}$ ways. We randomly chose 50 possible partitions and evaluated the agreement on each partition. On a remark, the original two blocks of 7 items were thoughtfully designed to be parallel. Here, the forms were scrambled for method evaluation, where the resulting partitions may no longer be comparable in specific content coverage, difficulty, and other item characteristics, despite measuring a common unidimensional trait of PSTRE.

Similar to the simulation study, the mean-squared deviation (MSE) with respect to $\hat{\theta}^{(r)}$,

$$MSE(\hat{\theta}^{(s)}) = \frac{1}{N} \sum_{i=1}^{N} (\hat{\theta}^{(s)} - \hat{\theta}^{(s)})^2,$$  \hspace{1cm} (2.5)
and the Kendall’s τ with \( \hat{\theta}_Y^{(r)} \) can be computed for each estimator produced from the scoring set.

Unlike the true \( \theta \), \( \hat{\theta}_Y^{(r)} \) is estimated based on final responses to only 7 items and contains measurement error. The correlation between \( \hat{\theta}_Y^{(s)} \) and \( \hat{\theta}_Y^{(r)} \) is hence attenuated by the reliability of \( \hat{\theta}_Y^{(r)} \), and the MSE of \( \hat{\theta}_Y^{(s)} \) with respect to \( \hat{\theta}_Y^{(r)} \) is expected to deviate from the MSE of \( \hat{\theta}_Y^{(s)} \) with respect to true \( \theta \). Rather than interpreting the two evaluation metrics as the recovery of true proficiency, they can instead be regarded as the split-half (\( B_s \) and \( B_r \)) agreement of latent trait estimates, or, alternatively, as the strength of association between \( \hat{\theta}_Y^{(s)} \) and performance on similar tasks (\( \hat{\theta}_Y^{(r)} \)). Lower MSE and higher Kendall’s τ hence suggest higher reliability.

On the scoring set, the response- and process-incorporated estimators were obtained following similar procedures as in the simulation studies. Specifically, the dimension of the process features was selected for each item by cross-validation, and the final score to each question was added as an additional process feature to guarantee perfect predictability. To evaluate performance under different test lengths, similar to experiment II, the number of items in \( B_s \) used for scoring ranged from 2 to 7. Specifically, when only two items were used for scoring, it was assumed that examinees’ processes and scores were observed only on the first two items. Subsequent items in the scoring set were added one by one until all 7 items were used. Because the final responses were polytomous, the generalized partial credit model described in Section 1.1.2 was used to calibrate the item parameters and to obtain the response-based \( \theta \) estimates. Additionally, for the second conditional expectation in Procedure 1, \( E(\hat{\theta}_Y | T_{X_{-j}}, Y_j) \), the polytomous final response \( Y_j \) was dummy-coded as regressors in the second regression.

**Results**

Figure 2.4 compares the MSE of the response- and process-incorporated trait estimators. Results are presented for different test lengths, ranging from 2 to 7 items, in the scoring set. The green boxes correspond to the response-based (baseline) \( \hat{\theta}_Y^{(s)} \), and the red boxes correspond to the process-incorporated \( \hat{\theta}_Y^{(s)} \). Each box plot in Figure 2.4 represents the distribution of the MSE across the 50 randomly sampled partitions of the 14 items into scoring and reference sets. One
can observe that for all test lengths, the process-incorporated latent trait estimator consistently demonstrated smaller MSE, indicating higher agreement with the performance on an external set of similar tasks (i.e., the reference set).

In particular, with two items, the process-incorporated estimator achieved comparable median MSE as the response-based estimator using five items. With four or more items, the process-incorporated $\hat{\theta}$ consistently achieved similar, if not lower, MSE than the response-based estimator using all 7 items.

![Figure 2.4: Response-based (baseline) and process-incorporated estimators’ MSE with respect to reference-set ability estimate on the PIAAC data.](image)

The box plots for the Kendall’s $\tau$ of the two types of estimators are presented in Figure 2.5. The correlations with reference set performance were consistently larger using process-incorporated scoring for all test lengths, suggesting that the rankings of latent ability estimates generated based on the problem-solving processes were more similar to the rankings on reference set performance. Again, scores based on processes required less items to achieve a given level of agreement. For instance, with 4 items, the process-incorporated estimator achieved similar or higher Kendall’s $\tau$
when compared to the response-based estimator with all 7 items. Attenuated by the reliability of
the reference set \( \theta \) estimate, the absolute Kendall’s \( \tau \) were mostly below .5. When compared to the
true \( \theta \), however, one would expect the correlation to be higher.

![Figure 2.5: Distribution of response-based (baseline) and process-incorporated Kendall’s \( \tau \) with
reference set performance across 50 partitions on the PIAAC data.]

2.4.2 Performance by Degree of Process- and Response-based Score Discrepancy

**Evaluation Methods**

The comparisons above focused on the overall agreement of each estimator with reference set
performance. One may also be interested in how the two methods perform for different types
of examinees. In particular, it is worth evaluating the relative performance of the two estimators
when they disagree on an examinee’s latent proficiency ranking. On each of the 50 partitions, we
computed the process-incorporated and response-based estimators using all 7 items on the scoring
set. We further regressed the response-based \( \hat{\theta}_Y^{(s)} \) on the process-incorporated \( \hat{\theta}_X^{(s)} \) using OLS and
computed each individual’s Studentized residual for the regression. Individuals were then binned
into 10 groups based on their deciles of the Studentized residuals. The deciles of the Studentized
residuals reflect the relative discrepancies in performance ranking based on the two trait estimators:

For individuals in the first decile, their performance rankings based on process were much lower than that based on responses. Individuals in the 10th decile, on the other hand, were ranked much higher based on responses than based on process. Individuals closer to the middle (4th - 6th decile) received similar rankings based on process and responses. The MSEs of the two trait estimators with respect to $\hat{\theta}_Y^{(r)}$ within each decile were then computed.

Results

The box plots of the MSEs with respect to reference set performance $\hat{\theta}_Y^{(r)}$ across the 50 partitions, separated by residual deciles, are shown in Figure 2.6. When the two scores agree on individuals’ rankings, the MSEs of $\hat{\theta}_X^{(s)}$ and $\hat{\theta}_X^{(s)}$ were similar. However, as we move towards the two ends where the two estimators started to disagree, the MSEs of the process-incorporated estimator were remarkably lower than that of the response-based estimator. Intuitively, the two estimators can be thought of as two judges, one judging individuals’ performance considering the problem-solving processes, and the other judging solely based on the final outcome. When the two judges disagree, the process-incorporated estimator consistently better predicted performance on similar tasks.

2.4.3 Empirical Interpretations of Process- and Response-based Score Discrepancy

Methods

The results above suggest that the proposed process-incorporated latent trait estimation procedures led to an increase in consistency with proficiency estimate on an external set of items, and that the improvement appeared most significant for individuals whose process-incorporated and response-based latent trait estimates disagreed most. One question worth asking is how the proposed approach scores individuals differently compared to the response-based counterpart. We explored this question by looking at the sequences of individuals whose process-incorporated and response-based latent trait estimates disagreed the most, that is, those with the highest or lowest
Studentized residuals for $\hat{\theta}_Y \sim \hat{\theta}_X$.

This time, with the purpose of interpretation rather than performance evaluation, all 14 items were used to obtain the two trait estimates. For the individuals in the bottom and top 10 of the Studentized residuals, we visually examined their action sequences on the 14 items.

**Results**

Figure 2.7 shows the scatter plot of each respondent’s $\hat{\theta}_X$ (x-axis) and $\hat{\theta}_Y$ (y-axis). Blue triangles correspond to the 10 individuals with the highest Studentized residuals, when regressing $\hat{\theta}_Y$ on $\hat{\theta}_X$.

These individuals received lower rankings based on processes than based on final responses.
For most of them, certain questions were successfully solved, but with less efficient strategies: For instance, to look for the requested information from a long spreadsheet, some of these examinees visually inspected every single entry, although a much more efficient strategy is to use “Search” or “Sort” (e.g., items U03a, U19a). To reply to an email sent to a group, some of them hand-typed the long list of email recipients, when they could simply press “Reply to all” or copy-and-paste (e.g., item U16). Aside from inefficient strategy usage, a few examinees also performed a large number of redundant steps, that is, actions that were not required for successful task completion.

The red rectangles in Figure 2.7, on the other hand, represent the 10 examinees with the lowest Studentized residuals. These examinees received higher ranking based on their problem-solving processes than based on final scores. Several common patterns were observed from their log data: The first was partial completion, where the examinee performed some of the key steps on a question, but, before reaching a credited response, proceeded to the next question by clicking “Next, Next_OK”. An example is item U16, which required sending an email to a list of recipients containing some key information. Several of the 10 examinees created the email and filled in the correct content and recipients, but they proceeded to the next question without clicking “Send”. Another common pattern was careless mistakes, where the examinee demonstrated the required skills for completing the task but slipped on an item due to carelessness. For example, on item U11b, which required sorting emails in the “Saved” folder, four of the ten examinees sorted the emails in the “Inbox”, the default, wrong folder. Intuitively, occasional carelessness and misinterpretation of question requirements, which lead to incorrect responses despite having the requisite skills, may be regarded as one of many sources of random measurement error. With additional information from the problem-solving processes incorporated, the proposed procedure for process-incorporated scoring appeared less impacted by such sources of measurement error.
Figure 2.7: Scatterplot of process-incorporated and response-based $\theta$ estimates with 14 items.

2.5 Discussions

2.5.1 Findings and Implications

Problem-solving processes contain rich information on individual characteristics, including the measured construct. The current study introduces a method to refine outcome-based latent trait estimates using the additional information from the problem-solving processes. A Rao-Blackwellization approach was proposed for the score refinement. Aside from choosing an appropriate IRT model for the final responses, the proposed approach is relatively data-driven and does not involve prior specification of a measurement model for the problem-solving processes, requiring less subjective inputs compared to expert-defined rubrics for process-incorporated scor-
The main theorem states that, under some regularity conditions, the proposed approach can lead to MSE reduction in latent trait estimation. Results from simulation studies corroborate the theorem. An empirical study using the PIAAC PSTRE data further showed that the process-incorporated latent trait estimate tended to have higher agreement with performance on similar tasks, thus higher reliability, compared to the response-based trait estimate. In addition, in order to achieve a particular level of reliability (i.e., MSE or $\tau$ with the external set of items), far fewer items would be required if the additional information from the problem-solving processes is exploited for scoring.

While the improvement in reliability by incorporating process information was consistent across test lengths, one particular merit of incorporating the process information into scoring lies in its use for short tests: For short tests, the final responses alone can rarely generate reliable trait estimates. Process information can thus be particularly useful for low-stakes computer-based assessment scenarios, when the administration of long tests is unrealistic or burdensome. With the additional information from problem-solving processes, the tests can be significantly shortened without compromising measurement reliability. An example is interim formative assessments during the learning process, where, after every one or few classes, instructors may want to learn how well the students have mastered the recently taught contents. Administration of a long test after every several classes can be very burdensome for the students and may interrupt the learning process. On the other hand, a relatively reliable latent ability estimate can be obtained if the problem-solving processes to a few constructed response items are available.

Although computerized adaptive testing [CAT; e.g., 44] can also reduce the required test length through the adaptive selection of test items tailored to individuals’ real-time proficiency estimates, the construction of a traditional CAT usually requires a large pre-calibrated item pool with hundreds of items, which may be overly costly and hard to achieve for many smaller-scale and low-stakes assessments. The production of a process-incorporated scoring rule, on the other hand, only requires sufficient items for reliable measurement of latent proficiency and a sample size that is
sufficient for item parameter calibration, process feature extraction, and training the regression models. Additionally, a process-incorporated CAT framework will be presented in Chapter 4.

2.5.2 Practical Considerations

The proposed approach could consistently improve test reliability. However, there are a few caveats to its implementation, especially for exams with higher stakes. First, in the empirical study, the performance of the process- and response-based latent trait estimators were evaluated using up to 7 items for scoring. The choice of up to 7 items was due to the limited number of total items available (14) and the need to set aside a large enough reference set of items used for evaluations. For an operational test, however, 7 items’ final responses are far from sufficient for reliable measurement, and the measurement error in the final response-based latent trait estimates can propagate to the process-incorporated scores through the conditional expectation. Test developers are advised to employ a sufficiently large scoring set so that reliable response-based latent trait estimates can be obtained. Besides, while we employed a random separation of the 14 items for evaluation, the original PSTRE test was designed in two parallel blocks with careful consideration of item difficulties and placements. As a result, our reported performance might not represent the optimal outcome achievable with a more thoughtful design.

Second, the proposed process-incorporated scoring approach aimed at improving the measurement precision, or reliability, of the assessment through MSE reduction. The validity of the scoring rule, however, is a separate critical issue that deserves further attention. Looking at the empirical interpretations of the process-incorporated scores, it appeared that individuals were scored higher based on processes when they gave up on the track to a correct response, demonstrated partially correct responses, or slipped on the final response due to careless mistakes. In these cases, increasing the individuals’ latent trait estimates may be reasonable, because each of these patterns constitutes evidence of partial or full mastery of the required skills for completing the tasks. Meanwhile, individuals who reached correct responses but with less efficient problem-solving strategies received lower process-incorporated scores. Assigning different trait estimates based on the choice
of test-taking strategies may be more controversial: On one hand, usage of more efficient problem-solving strategies was found positively correlated with the final score on other tasks in the empirical example, providing validity evidence on the use of such information for proficiency assessment. From this perspective, problem-solving strategy information may be incorporated into ability estimation similar to other types of collateral information, such as response times [e.g., 45, 24] and other covariates in latent regression [e.g., 46]. On the other hand, test takers may be unaware that they are scored based on information aside from task completion, raising concerns about the face validity and broader implications of the scoring criterion. At the same time, it raises test design questions such as whether examinees should be instructed that their scores can be affected by the problem-solving process. Evaluation of measurement validity would involve not only a search for empirical evidence that support the use and interpretation of the scores but also the appraisal of social consequences of score uses [47]. At the same time, scoring algorithms based on process data bring forth new validity and equity questions, such as whether different subgroups of test takers behave similarly and how to evaluate potential algorithmic bias [e.g., 34, 48]. We leave the discussion of how to best assist experts with the validation of data-driven latent proficiency estimators to Chapter 3.

Third, similar to item response modeling, the process-incorporated scoring rule should also be generalizable to the population. Conceptually, many high-dimensional regression methods search for a set of parameters that minimizes the expected loss at the population level instead of the finite sample of training data. In the presence of noise in the process data, this mandates an adequate sample size as well as variable selection to prevent overfitting. We recommend users adopt similar simulation studies as in the current study, by manipulating the number and prevalence of possible actions based on the actual item, to approximate the sample size requirement for implementing the procedures. Measures should also be taken to prevent overfitting when one extracts the sufficient statistic by fitting the high-dimensional regression model involving process features, for example, by using variable selection in conjunction with cross-validation.

Lastly, to establish the theoretical results on improved measurement efficiency, several assump-
tions were made in the current framework. One should note that, if some of the assumptions are violated, the efficiency results may be discounted, and several sources of bias may be introduced to the estimator. One of such instances is when the final score on an item is affected by systematic, construct-irrelevant variance. The presence of construct-irrelevant variance in the process alone does not constitute an assumption violation: We conducted a small-scale simulation study where the response process on each item was generated to depend on both $\theta$ and an additional, construct-irrelevant trait, $\eta \perp \theta$. As long as the final responses were independent of $\eta$, the process-incorporated proficiency estimator remained uncontaminated by $\eta$. Furthermore, we do expect that the proposed estimator improves upon the response-based estimator even beyond the three assumptions. Although practical diagnostic procedures are not yet available for verifying these assumptions by the data, in practice, practitioners could also perform an analysis as demonstrated in the Empirical Example section, to get a sense of the potential improvement in reliability by adopting the process-incorporated score.

2.5.3 Future Extensions

The methods for data-driven score refinement based on problem-solving processes can be extended in several ways. To start, while the current study provides one approach to increasing measurement reliability with process information, other methods, such as latent regression with process as covariates and confirmatory models with both response and process indicators, may be developed. Unlike the ordinal final outcomes which are designed to assess the measured trait, the problem-solving process data is high-dimensional and can contain substantial construct-irrelevant variance. For these parametric models, effective methods for variable selection will be needed to parse out the signal ($\theta$-related information) from the “noise”. Another potential extension of the process-incorporated scoring method is to diagnostic assessments [e.g., 49], where, instead of measuring individuals on the continuous proficiency continuum, the goal is to classify individuals into latent classes based on their mastery status of discrete skills. Lastly, we excluded observations with omissions in the current analysis, because missingness due to omissions often requires careful
treatment in both item calibration and scoring [e.g., 50, 51, 52], which is beyond the scope of the current study. Omissions can entail either good-faith attempts at a question or a lack of motivation, in which case, simply treating it as wrong introduces construct-irrelevant variance to final response scoring. At the same time, process data introduces new opportunities for disentangling different types of omissions. Future research can examine how process data can be incorporated into the scoring of tests in the presence of omissions.
Chapter 3: Differential Item Functioning Reduction via Process Data

3.1 Introduction

Establishing measurement invariance across groups is a critical aspect of research in educational and psychological testing. Ideally, items on a test are supposed to function the same way for all individuals, regardless of the groups they belong to [53]. In reality, however, differential item functioning (DIF) may occur when there are differences in the functioning of an item across groups, oftentimes based on demographic characteristics [54]. A variety of approaches have been proposed to identify DIF [55, 56, 57]. Theses methods can broadly be classified into two categories: non-parametric methods and parametric methods.

Among non-parametric approaches, one of the earliest and simplest methods for detecting DIF is leveraged on the three-way contingency tables, where each table is based only on respondents who received the same observed score. Previous research has established many popular methods for DIF detection, including Mantel–Haenszel [MH; 58, 59], standardization [60], and SIBTEST [61]. In the MH and standardization approaches, the observed total score is adopted to match examinees based on their overall performance. In SIBTEST, a valid submeasure is estimated first using anchor items, and the conditioning variable is then chosen as the estimated measure rather than the observed score. These non-parametric approaches have been shown to be effective in identifying DIF, but they do not account for the underlying latent traits of the examinees, which may be important in some cases.

Parametric methods for detecting DIF are typically based on item response theory (IRT). Generally speaking, these methods involve estimating two separate IRT models for the focal and reference groups, and then using comparative methods to test whether there is a significant difference between the groups. For example, this can be done using logistic regression with a likelihood ratio
test or chi-square test [62], or an IRT model with likelihood ratio test[63, 64]. Other methods include the item characteristic curve (ICC) area difference approach [65], or the H-statistic approach [66].

While the existing research has mainly focused on detecting and reporting the impact of DIF, items with significant DIF are typically discarded and excluded from operational use, despite the costs and efforts invested in item development and field testing. Hence, there has been a surging interest in identifying the causes that underlie the systematic differences in subgroup performances, and thereby eliminate target ability irrelevant influence [e.g., 67]. In the literature, several approaches have been proposed to identify the nuisance trait [e.g., 68, 69, 70] that leads to the group differences in item performance while controlling for target ability. However, when only final response data are available, it may be challenging to explain and model DIF without directly observing or measuring the predefined nuisance dimension.

This chapter aims to reduce or even remove the DIF among subgroups, and the underlying premise is that responses to items depend solely on cognitive processes, which are independent of demographic variables. In other words, two students with exactly the same cognitive process should have the same response distribution, regardless of which subgroup they belong to.

Observing comprehensive response processes brings new possibilities to alleviate DIF caused by several factors in our framework. Firstly, models, based on which assessments are made, are often overly simplified so that some important skills or attributes for the responses are ignored. Furthermore, such factors are unevenly distributed among subgroups. Thus, the observed empirical item response function is in fact the conditional distribution of responses given the attribute(s) of interest with the other (latent) factor integrated out. Consequently, this function will differ among subgroups.

In this chapter, we propose a method to systematically reduce DIF due to the presence of such nuisance attributes, even if they are not directly observed or accurately identified. To achieve this, we consider response processes as proxies to those nuisance attributes and include them in the measurement model. Process data contain rich information regarding each individual and are ideal
candidates to serve as proxies for a wide range of potential nuisance attributes.

However, the incorporation of process data may come at the cost of reduced measurement precision. This is due to the fact that process data include information not only about the nuisance attributes, but also about the target attribute. If too much information concerning the target attribute is included, the measurement model may lose its capability to accurately estimate the target attribute. On the other hand, from the DIF reduction perspective, it is desirable to incorporate as much process data information as possible to account for potential and unobserved nuisance attributes.

To address such issue, a variable selection mechanism is proposed to balance the inclusion of process data to account for the nuisance attributes and the maintenance of the estimability of the target attribute. If we are able to achieve a balance between these two goals, the measurement model and its scoring rule (e.g. maximum likelihood estimate of the target attribute) is free of DIF. Otherwise, we are only able to partially reduce DIF.

The rest of the chapter is organized as follows. The modeling framework and the process data feature selection method are proposed in Section 3.2. A simulation study is presented in Section 3.3. In Section 3.4, the proposed framework is applied to the data collected from 2012 PIAAC PSTRE. Concluding remarks are given in Section 3.5.

### 3.2 Differential Item Functioning Reduction

#### 3.2.1 Multidimensional Model for DIF

We present our method within a general multidimensional framework [71], where DIF is explained by the presence of multiple nuisance cognitive attributes in the item response function. Consider a scenario in which \( N \) individuals respond to \( J \) test items. Let \( Y_{ij} \) denote the response of person \( i \) to item \( j \) for \( i = 1, \ldots, N \) and \( j = 1, \ldots, J \). The entire sample is divided into two groups, with group membership denoted by \( g = r \) and \( f \), representing the reference group and the focal group, respectively.
Consider a general case that the item response function (IRF) for each individual $i$ is

$$p_j(\theta_i, \eta_i) \triangleq P(Y_{ij} = 1|\theta_i, \eta_i) \quad (3.1)$$

consisting of a multidimensional attribute vector $(\theta_i, \eta_i)$, where $\theta_i$ is the target attribute while $\eta_i$ is the nuisance attribute vector. Note that the IRF in Equation (3.1) holds for examinees in both the reference group and the focal group.

In the formulation in Equation (3.1), DIF arises when the conditional distribution $p(\eta_i|\theta_i, g = r)$ is different from $p(\eta_i|\theta_i, g = f)$. In this case, the marginal item response function for $\theta$ with $\eta$ integrated out is given by

$$p_j(\theta_i, g) = P(Y_{ij} = 1|\theta_i, g) = E[p_j(\theta_i, \eta_i)|\theta_i, g], \quad (3.2)$$

which differs for each group, e.g.

$$p_j(\theta_i, g = r) \neq p_j(\theta_i, g = f) \quad (3.3)$$

and thus, DIF occurs.

For instance, if $\eta_i|\theta_i, g = r$ is stochastically larger than $\eta_i|\theta_i, g = f$ in the usual stochastic order and the item response function $p_j(\theta_i, \eta_i)$ is an increasing function of $\eta_i$, then

$$p_j(\theta_i, g = r) = E[p_j(\theta_i, \eta_i)|\theta_i, g = r] > p_j(\theta_i, g = f) = E[p_j(\theta_i, \eta_i)|\theta_i, g = f], \quad (3.4)$$

for any $\theta_i$, resulting in uniform DIF. On the other hand, when the distribution of $\eta_i|\theta_i$ is group-invariant, $p_j(\theta_i, g)$ will be group-invariant and DIF does not exist for this item.

To summarize, in this setting, the presence of DIF can largely be attributed to three factors:

1. the item response function being dependent on both the target attribute $\theta$ and the nuisance attributes $\eta$; 2. the nuisance attributes $\eta$ having a heterogeneous distribution among the focal and
the reference groups; 3. the IRT model not including all the nuisance attributes in its item response function. The simultaneous presence of these three factors leads to differential item functioning.

3.2.2 DIF Reduction Model

As discussed in the previous section, the presence of unbalanced distributions of $\theta$ and $\eta$ among subgroups and missing $\eta$ from the measurement model causes DIF. To address this issue, a straightforward approach is to include the missing $\eta$ in the measurement model. However, it is generally difficult to systematically observe additional latent factors.

Process data, denoted by $S_{ij}$, are generated during the testing process and contain detailed information about an individual’s response process. These data can serve as proxies for a wide range of potential nuisance attributes. Specifically, process data refer to the log files generated by computer-based items that track all the observable actions, including mouse clicks and keystrokes, during the human-computer interaction process. Compared to the final response $Y_{ij}$, process data provide much more comprehensive information about students’ cognitive processes, making them a promising proxy for the additional latent variables $\eta$.

Another important property of process data is that the final response $Y_{ij}$ is a deterministic function of the responses process. This property is essential for removing DIF, as it ensures that the response $Y_{ij}$ depends only on the cognitive characteristics and thus, there exists an item response function as in Equation (3.1) valid for all subgroups. However, this IRF is of no use for assessment purposes, as it does not depend on the target attribute $\theta$ either because process responses contain information about both the target and the nuisance attributes. Nevertheless, this observation justifies the existence of at least one form of IRF that holds for all subpopulations, nonetheless can not be used directly. Further discussion on this topic is provided in Section 3.2.3.

Suppose the final responses are binary, i.e., $Y_{ij} = 1$ if the answer is correct and 0 otherwise. In this chapter, we consider the multidimensional two-parameter logistic (M2PL) model that is a
conventional multidimensional IRT model [61]. The item response function is

$$p_j(\theta_i, \eta_i) = P(Y_{ij} = 1|\theta_i, \eta_i) = \frac{e^{a_j\theta_i+a_j^T\eta_i+b_j}}{1 + e^{a_j\theta_i+a_j^T\eta_i+b_j}}, \quad (3.5)$$

where $a_j^T = (a_j, a_{j1}^T)$ and $b_j$ are the item slope and intercepts, respectively. We further write

$$\eta_{ij} = a_{j1}^T\eta_i \quad (3.6)$$

as the overall nuisance trait.

The process response $S_{ij}$ are typically difficult to include in measurement models directly as proxies to $\eta_{ij}$ because action sequences are in a non-standard format. Instead of working directly with $S_{ij}$, we extracted $K$-dimensional numerical features from $S_{ij}$, denoted $X_j = (X_{j1}, \ldots, X_{jk}) \in \mathbb{R}^K$. Various feature extraction methods can be used, such as MDS or sequence-to-sequence autoencoders as described in Section 1.3. These methods map each response process to a $K$-dimensional feature vector.

Afterwards, the extracted process features $X_{ij}$ can serve as a proxy of $\eta_{ij}$. In particular, we consider a simplified situation that

$$\eta_{ij} = \sum_{p=1}^P \omega_{jp}X_{ijp}. \quad (3.7)$$

The reader may consider more complex nonlinear models for their own analysis.

Therefore, the item response function is

$$P(Y_{ij} = 1|\theta_{ij}, X_{ij}) = \frac{e^{a_j\theta_{ij}+\sum_{p=1}^P \omega_{jp}X_{ijp}+b_j}}{1 + e^{a_j\theta_{ij}+\sum_{p=1}^P \omega_{jp}X_{ijp}+b_j}}. \quad (3.8)$$

The model structure is visualized in Figure 3.1.

 Compared to the classic 2PL model, e.g.,

$$P(Y_{ij} = 1|\theta_i) = \frac{e^{a_j\theta_i+b_j}}{1 + e^{a_j\theta_i+b_j}}, \quad (3.9)$$
the item response function in Equation (3.8) includes a correction term which is a function of corresponding process features, $X_{ij}$. It is parameterized by $\omega_j$, $a_j$, $b_j$ with inputs $\theta_i$ and $X_{ij}$.

If the observed process data features are able to account for the effect of the nuisance attributes, we expect that item response function Equation (3.8) is identical in the focal group and the reference group. We will empirically verify it in the subsequent real data analysis.

3.2.3 Process Feature Selection

Process data presents both an opportunity and a challenge in accounting for nuisance attributes. On the one hand, process responses contain comprehensive information about an individual’s cognitive and social characteristics, making them a promising proxy for the unobserved nuisance attributes. On the other hand, this comprehensive information also leads to potential identifiability issues in Equation (3.8). Specifically, when the binary response $Y_{ij}$ is highly predictable by the process data responses, the item response function Equation (3.8) may become close to a singular distribution, resulting in little to no information about the underlying target attribute $\theta$ [72].

To strike a balance, we need to carefully select a subset of process features to include in the model. Ideally, we want to include as many process features as possible to reduce DIF and account for nuisance attributes, while also limiting the amount of $X$ to ensure that the distribution of $Y$ remains away from singularity and there is enough information to estimate $\theta$. To address this challenge, we propose a variable selection method to identify a subset of $X$ that is just enough to reduce DIF and maintain the estimability of $\theta$. 

Figure 3.1: Example model structure without intercepts.
Consider the model in Equation (3.8) and let
\[ p_{rj}(\theta_i, X_{ij}) = P(Y_{ij} = 1|\theta_i, X_{ij}, g = r) \quad \text{and} \quad p_{fj}(\theta_i, X_{ij}) = P(Y_{ij} = 1|\theta_i, X_{ij}, g = f). \]

(3.10)

If DIF were completely removed, we expect that
\[ d_j^2 = \frac{1}{N} \sum_{i=1}^{N} (p_f(\theta_i, X_{ij}) - p_r(\theta_i, X_{ij}))^2 = 0, \]

(3.11)

that is, the item response function including appropriate process data features is identical for both the reference group and the focal group.

If some items in the test exhibit no DIF, the target attribute \( \theta_i \) could be estimated with reasonable accuracy. In the variable selection, we assume that \( \theta_i \) is known, and it is estimated using responses to other items in practice. Let \( \tilde{\theta}_i \) be the estimated target attribute based on DIF-free items’ responses. We fit the model in Equation (3.8) separately for both the reference and the focal groups, and obtain estimated item response functions \( \hat{p}_{fj}(\tilde{\theta}_i, X_{ij}) \) and \( \hat{p}_{rj}(\tilde{\theta}_i, X_{ij}) \). We then compute their observed \( L_2 \) distance as in Equation (3.11). Empirically, this amounts to
\[ \hat{d}_j^2 = \frac{1}{N} \sum_{i=1}^{N} [\hat{p}_{fj}(\tilde{\theta}_i, X_{ij}) - \hat{p}_{rj}(\tilde{\theta}_i, X_{ij})]^2. \]

If DIF does not exist in this item, \( \hat{d}_j \) should be substantially close the zero. Therefore, we select a subset of \( X \) to minimize \( \hat{d}_j \) such that the DIF is insignificant and make sure the item response function is still \( \theta \)-relevant.

It is worth mentioning that the value of \( \hat{d}_j \) does not necessarily decrease monotonically with the increasing number of process features included in the model. If the added features are irrelevant to the nuisance attributes, they will contribute to the variance of \( \hat{p} \) and thus, will increase the value of \( \hat{d}_j \). Later in our analysis, we will observe that the behavior of \( \hat{d}_j \) as a function of the number of process features is characterized by an initial sharp decrease, followed by a gradual increase as more irrelevant features are added to the model.
The following theorem provides a reference distribution of $\hat{d}_j$ to quantitatively assess the magnitude of $L_2$ distance.

**Theorem 3.** Let subscripts $r$ and $f$ indicate the parameters in Equation (3.8) for the reference group and the focal group. Suppose that $a_{fj} = a_{rj} = a_0, b_{fj} = b_{rj} = b_0, \omega_{fj} = \omega_{rj} = \omega_0$, then $N \hat{d}_j^2$ converges in distribution to a generalized chi-square distribution as $N \to \infty$, i.e.,

$$N \hat{d}_j^2 \xrightarrow{d} \xi = \sum_{i=1}^{K} \lambda_i Z_i^2$$  \hspace{1cm} (3.12)

where $Z_1, \ldots, Z_K$ are independent, standard normal variables and $\lambda_i$s depends on $a_0, b_0, \omega_0$, the joint density function of $\theta_i$ and process features $X_{ij}$, and the limit of the ratio between focal and reference group size. In addition, $\lambda_i$s can be estimated consistently from the data.

The proof of Theorem 3 is provided in the appendix. The cumulative distribution function of the generalized chi-square distribution does not have an explicit form, so we compute its expectations and quantiles by Monte Carlo methods.

If $\hat{d}_j$ corresponding to the selected set of $X$ falls within a certain range of the above reference distribution (e.g. $p$-value), we expect that DIF is largely removed. Otherwise, the effect of the nuisance parameters is not completely accounted for by process data features, and only a certain reduction in DIF is achieved.

### 3.2.4 Target Latent Trait Estimation

After selecting the subset of process features, we pool the data from both the reference group and the focal group examinees to estimate the parameters $\omega_j, a_j, b_j$ in Equation (3.8). The target latent attribute is re-estimated for each individual using the updated IRF. For items with process features added in their item response functions, the updated IRF becomes
\[
P(Y_{ij} = 1|\theta_i, X_{ij}, \omega_j, \hat{a}_j, \hat{b}_j) = \frac{e^{\hat{a}_j \theta_i + \hat{b}_j}}{1 + e^{\hat{a}_j \theta_i + \hat{b}_j}} = \frac{e^{\hat{a}_j \theta_i + \sum_{p=1}^{P} \hat{\omega}_{jp} X_{ijp} + \hat{b}_j}}{1 + e^{\hat{a}_j \theta_i + \sum_{p=1}^{P} \hat{\omega}_{jp} X_{ijp} + \hat{b}_j}}, \quad \text{for} \quad j = 1, \ldots, J. \quad (3.13)
\]

If maximum likelihood estimation is adopted to estimate target ability, the updated estimate of \( \theta_i \) with DIF reduction, denoted \( \hat{\theta}_i \), will be given by

\[
\hat{\theta}_i = \arg \max_{\theta} \prod_{j=1}^{J} P(Y_{ij} = y_{ij} | \theta, X). \quad (3.14)
\]

3.2.5 Procedure

A practical roadmap is outlined as follows for a set of \( J \) items administered to \( I \) subjects, where the item responses are assumed to follow a 2PL model. Suppose that final response DIF has been detected on a subset of items, \( B \), using a DIF detection method (e.g., \( L_2 \) distance method in Theorem 3, likelihood ratio test). The rest of the items, \( B^C \), are treated as DIF-free anchor items, for which group-invariant 2PL item parameters, \( \hat{a}_j, \hat{b}_j \), have been estimated for each \( j \in \{B^C\} \).

The steps below implements the proposed DIF reduction procedure by means of process features.

1. Obtain initial latent trait estimate, \( \tilde{\theta}_i \), for each person \( i \) using the anchor items without DIF.

2. For each item \( j \) in \( B \), perform the forward stepwise logistic regression method to estimate nuisance trait \( \omega_j \) and DIF-corrected IRT model parameters, \( a_j, b_j \). Specifically:

   (a) Model initialization: With the initial ability estimates, \( \tilde{\theta}_i \), fit logistic regression model, \( Y_{ij} \sim \tilde{\theta}_i \), separately for the focal and reference groups. Calculate the observed \( \hat{d}_j \) between the reference and focal groups and compute the \( p \)-value for testing the hypothesis \( H_0 \) that DIF is significant.
(b) Step-wise feature selection: For each of the remaining process features (i.e., candidate features) that has not yet been added to the current model:

i. Add the feature to the current model, and perform logistic regression on focal group and reference group separately.

ii. Check whether the coefficients of the added feature and $\tilde{\theta}_i$ are significant.

iii. Calculate $\hat{d}_j$ with the feature added to the current model.

The feature that achieves the lowest $\hat{d}_j$ and satisfies (b)(ii) will be added to the current model. Update the current model, and perform hypothesis testing and calculate the corresponding $p$-value by the asymptotic distribution.

(c) Repeat step (b) until either DIF is substantially removed or no remaining process features can be added to satisfy (b)(ii).

(d) Final model selection: Select the dimension of process features, $k^*$, to be the smallest one that results in the empirical $L^2$ distance $\hat{d}_j$ being significantly close to 0. The final item response function contains $\tilde{\theta}_i$ and the first $k^*$ selected process features (denoted $X^*_ij$).

(e) IRF update and offset calculation: Using the entire sample (focal and reference), refit the logistic regression $Y_{ij} \sim \tilde{\theta}_i + X^*_ij$. Set $\hat{a}_j$ to be the coefficient in front of $\tilde{\theta}_i$ and $\hat{\omega}_j$ to be the coefficient vector of $X^*_ij$.

3. Final target trait estimation: For each person $i$, using responses to all $J$ items, re-estimate target ability $\theta$ with the updated IRFs. A maximum likelihood estimate $\hat{\theta}_i$ is obtained as in Section 3.2.4, where

$$P(Y_{ij} = y_{ij} \mid \theta) = \frac{e^{\hat{a}_j \theta + X^*_ij^T \hat{\omega}_j}}{1 + e^{\hat{a}_j \theta + X^*_ij^T \hat{\omega}_j}}$$

for $j \in B$, and

$$P(Y_{ij} = y_{ij} \mid \theta) = \frac{e^{\hat{\alpha}_j \theta + b_j}}{1 + e^{\hat{\alpha}_j \theta + b_j}}$$

for $j \in B^C$, with $\hat{\alpha}_j, \hat{b}_j$ as the original 2PL IRT parameter estimates before DIF reduction.
3.3 Simulation Studies

In this section, the proposed framework was applied to simulated data sets, where items were designed to exhibit DIF. The purpose is to demonstrate that, in the presence of DIF, the proposed framework could result in better latent trait estimates compared with traditional IRT model. We simulate data with $I = 10000$ examinees and $J = 20$ items. Half of these items were simulated to exhibit DIF, while the remaining serve as anchor items. Among all examinees, one-fourth were randomly selected to be from the focal group, and the remaining were from the reference group. Two scenarios of DIF were considered.

3.3.1 Uniform DIF

In the uniform DIF settings, for each examinee, the true target ability was generated from the standard normal distribution, i.e., $\theta_i \sim N(0, 1)$. For each item, the 10-dimensional process features were generated from a multivariate normal distribution. To be more specific:

$$X_{ij} \sim N(\mu_{gj}, I_p), \quad (3.15)$$

where each element of $\mu_{fj} \sim \text{Unif}(-0.5, 0)$, each element of $\mu_{rj} \sim \text{Unif}(0, 0.5)$, and $I_p$ is the $p \times p$ identity matrix. The nuisance latent ability was assumed as a linear combination of process features, i.e., $\eta_{ij} = \omega_j^T X_{ij}$ where

$$\omega_{jp} \sim \text{Exp}(1), \text{ for } p = 1, \ldots, 10.$$

Let $a_j = 1$ and $b_j \sim \text{Unif}(-1, 1)$ for $j = 1, \ldots, J$. The item responses were then simulated based on the nuisance model in Equation (3.8). Note that all coefficients were simulated to be positive. As a result, the nuisance ability of the reference group was stochastically larger than that of the focal group, which guarantees the existence of uniform DIF.

The simulation was replicated 50 times. For each replication, examinees’ target trait, $\theta$, was
estimated based on either the regular 2PL model or based on the DIF-reduction framework. Specifically, under the regular 2PL model, items were calibrated assuming the typical unidimensional 2PL IRT model given $\theta$, and individuals’ latent traits estimates, $\hat{\theta}_{i}^{IRT}$, were then obtained based on the final responses. Under the proposed DIF-reduction framework, the IRT estimate based on anchor items was treated as the initial estimate $\tilde{\theta}_{i}$. The proposed procedures for DIF reduction was then applied to items exhibiting significant DIF. Final latent trait estimates, $\hat{\theta}_{i}$, were derived based on the updated IRFs incorporating the process features.

Two methods were compared in terms of the root mean square error (RMSE) and Pearson correlation between true and estimated target latent traits. Figure 3.2 displays the scatterplots of the RMSE and Pearson correlation with true $\theta$, respectively, using either the regular 2PL-based ($x$-axis) or the DIF-corrected ($y$-axis) target trait estimates in each of the 50 replications. The proposed procedure improved the recovery of individual target proficiency consistently by taking the nuisance latent trait into consideration.

3.3.2 Non-uniform DIF

In this section, a similar simulation setting was adopted. The only difference was that the target ability $\theta$ was simulated such that it might be correlated with the process features $X_{ij}$. To be more specific,

$$
\begin{pmatrix}
\theta \\
X_{ij}
\end{pmatrix} \sim N(0, \begin{pmatrix}
1 & 0.25 & \ldots & 0.25 \\
0.25 & 1 & 0 & \quad \\
\vdots & & \ddots & \quad \\
0.25 & 0 & \quad & 1
\end{pmatrix})
$$

(3.16)
Figure 3.2: RMSE and correlation between true ability and estimation with or without DIF reduction under uniform DIF.
Figure 3.3: RMSE and correlation between true ability and estimation with or without DIF reduction under non-uniform DIF.
for reference group and

\[
\begin{pmatrix}
\theta \\
X_{ij}
\end{pmatrix} \sim N(0, \begin{pmatrix}
1 & 0 & \ldots & 0 \\
0 & 1 & 0 & \ddots \\
\vdots & \ddots & \ddots & \ddots \\
0 & 0 & \ldots & 1
\end{pmatrix})
\] (3.17)

for focal group. The difference of the distribution of \((\theta, X_{ij})\) guarantees the existence of non-uniform DIF in the simulation.

The simulation was repeated 50 times, and the \(\theta\) recovery results with the two methods are shown in Figure 3.3. The results illustrate the superior performance of the proposed framework in reducing the impact of DIF on the latent trait estimates.

### 3.4 Empirical Example: PIAAC PSTRE

#### 3.4.1 The PIAAC PSTRE Data

The proposed DIF reduction framework was applied to the process and response data from the PIAAC PSTRE data. The current study used the process and response data on \(J = 13\) PSTRE items \(^1\). The sample was chosen as the \(I = 12501\) individuals from 17 countries and regions who responded to all 13 items. For the sake of the current study, polytomous final responses were recorded into dichotomous responses, based on the criterion that binary response as 1 for polytomous score 2 or 3; binary response as 0 for polytomous score 0 or 1. There were 135 examinees who answered all questions correctly and 947 who got all 0s.

Selected descriptive statistics for each of the 13 items are displayed in Table 3.1, where \(N\) denotes the number of possible actions, \(\bar{L}\) is the average process length, and Correct % is the percentage of correct responses. Note that due to differences in the population under consideration, some summary statistics presented here differ from those presented in Table 1.1. For each item, the problem-solving interface was composed of one or more simulated ICT environments, including

---

\(^1\)A total of 14 items were administered. One of the items was not included in the current study because of log data recording issues.
email client, spreadsheet and web browser. Some items such as U01a and U01b contained only one environment while other items, such as U02 and U23, involved multiple environments. The items also varied in the length and complexity of the processes, with number of possible actions ranging between 47 and 1659, and average sequence length ranging between 10.6 and 99.3 actions.

For many items, successful task completion required an intricate combination of 40 or more actions, which could be broken down into multiple key steps. In addition, for the same task, multiple strategies may exist to perform the same step. Therefore, while all items were designed to measure the unidimensional latent PSTRE ability, the complexity in the problem-solving processes makes it reasonable to consider the effect of nuisance traits on successful task completion.

The grouping variable considered in the current illustration was examinee age. The age of the respondents ranged widely from 16 to 65 as shown in Figure 3.4. Although more senior populations may be less familiar with ICT tools, it is also possible that certain design aspects of an item makes it more difficult for a senior test-taker to fully engage in. In these cases, nuisance traits may exist, making more senior individuals less likely to complete certain questions correctly compared to

Table 3.1: Descriptive statistics of the 13 PIAAC problem-solving items.

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
<th>N</th>
<th>$\bar{L}$</th>
<th>Correct %</th>
</tr>
</thead>
<tbody>
<tr>
<td>U01a</td>
<td>Party Invitations</td>
<td>234</td>
<td>24.3</td>
<td>70.6</td>
</tr>
<tr>
<td>U01b</td>
<td>Party Invitations</td>
<td>270</td>
<td>52.9</td>
<td>50.7</td>
</tr>
<tr>
<td>U02</td>
<td>Meeting Room</td>
<td>361</td>
<td>63.4</td>
<td>29.2</td>
</tr>
<tr>
<td>U03a</td>
<td>CD Tally</td>
<td>387</td>
<td>14.7</td>
<td>42.1</td>
</tr>
<tr>
<td>U04a</td>
<td>Class Attendance</td>
<td>1659</td>
<td>48.2</td>
<td>20.2</td>
</tr>
<tr>
<td>U06a</td>
<td>Sprained Ankle</td>
<td>47</td>
<td>10.6</td>
<td>24.9</td>
</tr>
<tr>
<td>U06b</td>
<td>Sprained Ankle</td>
<td>127</td>
<td>16.0</td>
<td>49.6</td>
</tr>
<tr>
<td>U07</td>
<td>Book Order</td>
<td>156</td>
<td>19.9</td>
<td>49.8</td>
</tr>
<tr>
<td>U11b</td>
<td>Locate E-mail</td>
<td>160</td>
<td>32.1</td>
<td>38.3</td>
</tr>
<tr>
<td>U16</td>
<td>Reply All</td>
<td>271</td>
<td>99.3</td>
<td>62.1</td>
</tr>
<tr>
<td>U19a</td>
<td>Club Membership</td>
<td>530</td>
<td>28.4</td>
<td>73.2</td>
</tr>
<tr>
<td>U19b</td>
<td>Club Membership</td>
<td>580</td>
<td>22.8</td>
<td>50.7</td>
</tr>
<tr>
<td>U23</td>
<td>Lamp Return</td>
<td>337</td>
<td>32.5</td>
<td>51.2</td>
</tr>
</tbody>
</table>

Note: $N$ = number of possible actions; $\bar{L}$ = average sequence length; Correct % = percentage of correct responses.
their younger counterparts with the same PSTRE ability. In Section 3.4.3 below, it is shown that final response DIF was observed on multiple questions when age was the grouping variable.

Figure 3.4: Distribution of examinee self-reported age for the PSTRE assessment data.

3.4.2 Process Features

The process features obtained via MDS can preserve a significant proportion of information in the original action sequences. Note that, for the PSTRE assessment, an examinee’s final outcome on a question is perfectly predictable from the recorded problem-solving processes. As shown in Table 3.2, using the MDS features of each item as predictors, the out-of-sample prediction accuracy of final binary response on each item exceeded 94% using logistic regressions and exceeded 96% using a one-hidden-layer neural network. While the optimal dimension $p$ of process features can be chosen item-by-item using cross-validation [18], for the current study, $p = 100$ was adopted to maintain the consistency and to ensure adequate amount of process information have
been preserved for each item.

Table 3.2: Out-of-sample accuracy for binary response prediction using MDS features.

<table>
<thead>
<tr>
<th>ID</th>
<th>Logistic Regression Accuracy %</th>
<th>Neural Network Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>U01a</td>
<td>96.2</td>
<td>97.4</td>
</tr>
<tr>
<td>U01b</td>
<td>93.6</td>
<td>94.2</td>
</tr>
<tr>
<td>U02</td>
<td>94.5</td>
<td>95.7</td>
</tr>
<tr>
<td>U03a</td>
<td>99.1</td>
<td>99.8</td>
</tr>
<tr>
<td>U04a</td>
<td>94.8</td>
<td>96.1</td>
</tr>
<tr>
<td>U06a</td>
<td>98.7</td>
<td>98.5</td>
</tr>
<tr>
<td>U06b</td>
<td>99.2</td>
<td>99.6</td>
</tr>
<tr>
<td>U07</td>
<td>99.9</td>
<td>100</td>
</tr>
<tr>
<td>U11b</td>
<td>96.5</td>
<td>96.7</td>
</tr>
<tr>
<td>U16</td>
<td>97.8</td>
<td>98.2</td>
</tr>
<tr>
<td>U19a</td>
<td>97.7</td>
<td>98.3</td>
</tr>
<tr>
<td>U19b</td>
<td>99.2</td>
<td>99.3</td>
</tr>
<tr>
<td>U23</td>
<td>99.4</td>
<td>99.5</td>
</tr>
</tbody>
</table>

Note: 4 hidden nodes and ReLU activation function are used in the neural network.

Recall that in Section 3.4.1, engagement was a possible nuisance ability which may vary from younger to elder. An inattentive respondent usually tried to skip a task directly or randomly clicked/typed, while an attentive examinee was more likely to explore more buttons and check answers before enter the next item. As a result, attentiveness in response process can be reflected in the process length. Figure 3.5 presents the correlation between first principal feature and the log-arithm of the process length. For most items, the correlation was higher than 0.8, which suggests process features as a suitable predictors for engagement.

3.4.3 DIF Existence

Following the discussion above, the focal group was chosen as elder respondents, with an age threshold of 48 (3rd quartile). This amounts to 3019 respondents in the focal group and 9482 in the reference group. Various conventional DIF detection methods are employed which suggest that item U02 and item U23 are DIF-free items while DIF effects were identified on the rest of the items to varying degrees. The initial estimate \( \tilde{\theta} \) was determined using the IRT estimate derived
Figure 3.5: Absolute correlation between first principal feature and logarithm of sequence length.

from the two DIF-free items. Another common approach to detect and report the effect of DIF is to fit a logistic regression with group interaction, that is,

\[ P(Y_i = 1| \theta_i, g) = \frac{e^{\tau_0 + \tau_1 \theta_i + \tau_2 g + \tau_3 \theta_i g}}{1 + e^{\tau_0 + \tau_1 \theta_i + \tau_2 g + \tau_3 \theta_i g}}, \]  

(3.18)

where an anchor-item-based estimate \( \tilde{\theta}_i \) is used in replacement of true ability \( \theta_i \), and \( \tau_2 \) and \( \tau_3 \) correspond to the group differences in item intercept and slope, respectively. Significant \( \tau_2 \) and \( \tau_3 \) values would suggest uniform and non-uniform DIF, respectively. Subsequently, hypothesis testing such as Wald tests can be conducted to examine whether \( \tau_2, \tau_3 \) significantly deviated from 0.

The observed \( L_2 \) distance statistic under the base model, that is the nuisance model without adding any process features, can also be used for DIF detection. This can be done by obtaining the \( p \)-value based on the asymptotic null distribution in Theorem 3. It is expected that the \( L_2 \)
distance-based test can detect both uniform and nonuniform DIF. The results for DIF detection, obtained from both approaches, are summarized in Table 3.3.

Table 3.3: DIF detection results using logistic regression and $L_2$ distance method.

<table>
<thead>
<tr>
<th>ID</th>
<th>$\hat{\tau}_2(\hat{\sigma}(\tau_2))$</th>
<th>$\hat{\tau}_3(\hat{\sigma}(\tau_3))$</th>
<th>$L_2$ Distance</th>
<th>$L_2$ Distance P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>U01a</td>
<td>-1.08(0.06)</td>
<td>-0.11(0.11)</td>
<td>0.200</td>
<td>0.00</td>
</tr>
<tr>
<td>U01b</td>
<td>-0.81(0.05)</td>
<td>0.43(0.10)</td>
<td>0.168</td>
<td>0.00</td>
</tr>
<tr>
<td>U03a</td>
<td>-0.75(0.05)</td>
<td>0.13(0.09)</td>
<td>0.146</td>
<td>0.00</td>
</tr>
<tr>
<td>U04a</td>
<td>-0.20(0.06)</td>
<td>0.35(0.11)</td>
<td>0.025</td>
<td>0.05</td>
</tr>
<tr>
<td>U06a</td>
<td>-0.14(0.06)</td>
<td>0.08(0.10)</td>
<td>0.020</td>
<td>0.14</td>
</tr>
<tr>
<td>U06b</td>
<td>-0.06(0.05)</td>
<td>0.31(0.08)</td>
<td>0.042</td>
<td>0.00</td>
</tr>
<tr>
<td>U07</td>
<td>-0.10(0.05)</td>
<td>0.27(0.10)</td>
<td>0.036</td>
<td>0.00</td>
</tr>
<tr>
<td>U11b</td>
<td>-0.70(0.05)</td>
<td>0.19(0.10)</td>
<td>0.130</td>
<td>0.00</td>
</tr>
<tr>
<td>U16</td>
<td>-0.61(0.05)</td>
<td>0.00(0.10)</td>
<td>0.120</td>
<td>0.00</td>
</tr>
<tr>
<td>U19a</td>
<td>-0.19(0.07)</td>
<td>0.32(0.12)</td>
<td>0.063</td>
<td>0.00</td>
</tr>
<tr>
<td>U19b</td>
<td>-0.34(0.05)</td>
<td>0.29(0.10)</td>
<td>0.077</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Note: Results in bold indicate statistical significance at .05 level.

The results were consistent across the two DIF detection methods, as the items with significant group slope or intercept effects in the multiple-group IRT method were also flagged with significant $L_2$ distances in the IRFs.

3.4.4 DIF Reduction

The proposed DIF procedures were applied to the 11 items detected with significant DIF. In each step of step-wise feature selection, in addition to the observed $L_2$ distance between the focal and reference groups given the corresponding set of features (Equation (3.11)), the threshold value of $L_2$ distance that corresponds to a significant level of 0.05 were also calculated. The results of the observed $L_2$ distance and the thresholds for hypothesis testing are presented in Figure 3.6, with each subplot displaying the results for one of the 11 items. The $x$–axis represents the number of features included in the nuisance model, and the two curves in yellow and blue represent the observed $L_2$ distance and the $L_2$ distance threshold for rejecting the null hypothesis that DIF does not exist in this item, respectively.
Figure 3.6: Observed $L^2$ distance and the hypothesis testing thresholds as a function of feature dimension.

It can be observed that, for each item, the observed $L_2$ distance (yellow) rapidly fell below the hypothesis testing threshold (blue) after incorporating several process features, suggesting that the focal and reference group IRFs no longer differed. However, if one kept adding additional process features, the observed $L_2$ distance slowly increased. Furthermore, the maximum, 0.90 quantile and median values of item response function distances evaluated at all people are shown in Figure 3.7. Although the maximal distance grew slowly as more process features were included, the median decreased rapidly, indicating that the response functions for focal and reference groups did converge to the same functional form for most respondents.
Figure 3.7: Maximum, 0.9 quantile and median of response function distances in log scale.

The Fisher information of parameter $a_j$ is plotted against the added feature dimensions in Figure 3.8. The less the Fisher information for parameter $a_j$, the less confidence we have for the estimation $\hat{a}_j$, which will affect the accuracy of final target trait estimation. The gradual decrease of Fisher information for both groups in Figure 3.8 validates the importance of variable selection.
Figure 3.8: Fisher information of parameter $a_j$ in log scale.

Besides, Figure 3.9 shows the sample mean of the Fisher information of the target ability $\theta$, as more process features are included for each item. For the plot of each item, the process features dimensions for the other items are fixed as the results from step-wise selection. It could be observed that the Fisher information for $\theta$ decreases as more process features are added for DIF reduction. Thus, if no additional constraints are imposed on the step-wise feature selection, adding too many process features for DIF reduction could result in an increase in the standard error of $\theta$ estimation, thus lower test reliability.
Figure 3.9: Sample mean of Fisher information of $\theta$ in log scale.

The nuisance traits constructed from the DIF reduction procedures contain valuable process information, which may be directly tied to empirical interpretations. As an example, in item U01a, respondents were asked to classify several emails in the inbox into different folders, based on the content of each email. Different strategies could be applied to move the emails, including using a toolbar icon, using the dropdown menu, or directly dragging an email and dropping it into the target folder. Among these, directly dragging and dropping emails proves to be the most efficient way, since it involves least actions and avoids mistaken clicks. 91.5% of the respondents using the drag-and-drop strategy partially completed the whole task, while 84.4% of the individuals were correct employing the toolbar and 87.0% using the dropdown menu. Interestingly, it was also noted that
the drag-and-drop strategy, despite being more efficient for moving emails, was significantly less favored within the focal group. Merely 52.0% of the elder subgroup employed this method, in stark contrast to 74.7% in the younger subgroup. The uneven application of the solving strategy consequently might lead to the presence of DIF.

Moreover, the empirical distributions of the nuisance trait from U01a, categorized by whether or not the respondent used the drag-and-drop method, is exhibited in Figure 3.10. It is clear that the nuisance latent trait identified through the DIF reduction procedures correlated with the choice of email-moving strategy and could be interpreted as the familiarity with the drag-and-drop method. Therefore, it is anticipated that the distribution of nuisance trait estimates varies significantly between the focal group and the reference group, as demonstrated in Figure 3.11. The results align with our observation that older respondents are less familiar with the drag-and-drop strategy.

The cause behind this age difference in strategy use is unclear: It is possible that senior individuals might be simply less familiar with this method for moving emails. Another possible interpretation is that, due to the small font size in the email client environment and the narrow gaps between consecutive emails or folders, senior individuals may find it more burdensome and
error-prone to move emails using drag-and-drop. Similar patterns were found on other items with detected DIF, where more senior individuals were less likely to employ problem-solving strategies that involve inspecting or exploring tables or lists with small font size and narrow line spacing, such as long spreadsheets.

For individuals with lower estimated nuisance trait, the expected increase in their correct target trait ($\theta$) estimate would be larger. Thus, for U01a, where the focal group tended to have lower nuisance trait values, the increase in the final estimation is expected to be larger than that for the reference group.

Last but not the least, the final latent ability estimates ($\hat{\theta}_i$) after DIF reduction are plotted against the 2PL IRT-based estimates using all items ($\hat{\theta}_i^{IRT}$) in Figure 3.12. Blue triangles correspond to the 10 individuals with the largest differences between $\hat{\theta}_i$ on $\hat{\theta}_i^{IRT}$. These individuals received lower rankings based on the DIF reduction model than the IRT model. Interestingly, all of them failed to complete the tasks at DIF-free items (item U02 and item U23) while their performance on the remaining items was relatively good. Besides, most respondents belong to the younger age group and their performance in handling the problems seems more influenced by their nuisance ability than their target ability. For instance, every individual in this group who was credited for item

![Graph showing empirical distribution and mean of nuisance ability estimation for both groups.](image-url)

Figure 3.11: Empirical distribution and mean of nuisance ability estimation for both groups.
U01a used the drag-and-drop strategy rather than the others.

In contrast, the yellow squares in Figure 2.7 represent the 10 examinees showing the lowest differences between $\hat{\theta}_i$ and $\hat{\theta}_i^{IRT}$. Notably, more of them belong to the focal group. Unlike the previously discussed group, all these individuals performed well on DIF-free items but fell short in their performance on the remaining items. For instance, in the case of item U01a, some of these individuals opted for the toolbar strategy but failed to successfully complete the task.

![Scatterplot of the IRT-based latent trait estimates \(x\)-axis and the DIF-corrected latent trait estimates \(y\)-axis.](image)

Figure 3.12: Scatterplot of the IRT-based latent trait estimates \(x\)-axis and the DIF-corrected latent trait estimates \(y\)-axis.
3.5 Discussion

This article presented a method to “debias” items detected with DIF using problem-solving process information. The method was established on the multidimensional 2PL IRT model for DIF, where group differences in item response functions can be explained by the presence of construct-irrelevant nuisance traits. A step-wise feature selection method was proposed, and the results with and without DIF reduction were evaluated both on simulated data and on the empirical data from the PIAAC PSTRE assessment. The simulation results demonstrated that the DIF reduction procedures could improve the estimation accuracy of the target latent trait. For the empirical example, based on the $L^2$-loss-based statistical hypothesis testing, it was observed that that age-related DIF existed among 11 of the 13 PSTRE items. After correcting for the DIF effects by incorporating process-based nuisance latent traits, individuals’ target latent trait estimate rankings have changed.

Item development is usually associated with high costs, as it involves a significant amount of effort from initial item writing to item reviews, field testing, and psychometric analysis. The cost for item development would be even higher for computer-based interactive items, which involves the additional work for interactive item interface development, log data storage, and feature engineering. When response DIF is detected, the conventional approach is to discard the item. With the proposed procedures, an item with DIF may be debiased by incorporating the nuisance trait into the IRT, making it possible to keep the item for operational tests, thus using resources more efficiently. Furthermore, the process-based nuisance traits can be tied to empirical interpretations. This information can help test developers understand the reasons behind detected DIF, providing guidance on the future development of assessments towards fairness and equity.

The current study has several limitations. First, the current procedures for step-wise feature selection considers only one objective, that is, the reduction in the $L_2$ distance in IRFs between the focal and reference groups. As can be observed from the results on the $\theta$ Fisher information, introducing excessive amount of process features may result in declines of test reliability. Future
research may look into feature selection methods that simultaneously balance multiple objectives, such as DIF reduction and preservation of target trait information.

Another limitation of the current study is the procedures that were implemented for detection of DIF, that is, the identification of the set of items that required DIF reduction. For simplicity, DIF detection for each item was performed by assuming all other items were DIF-free, that is, all other items were used as the anchor set. This approach may be prone to false alarms when a large proportion of items in the test exhibits DIF. Prior research on DIF detection have looked into alternative approaches for anchor set identification [e.g., 73, 74, 75], which could result more accurate identification of DIF and DIF-free items. In practice, these more sophisticated methods for scale purification and DIF detection may be employed to identify the set of items that require DIF reduction.

There are several possible extensions to the current approach. First, as increasingly many assessments are designed to be computer-based, time stamps of actions are recorded alongside the action sequences. Intuitively, the time elapsed between two consecutive actions may reveal additional information on top of the action sequences. Future studies can look into the extra information that inter-event times can provide on the modelling of nuisance traits.

Another possible extension is to consider nonlinear models for the process-based nuisance traits. For the current study, latent process features were extracted using MDS from the process data, and a linear model was used for the nuisance trait. This assumes that the nuisance trait is a linear combination of the MDS features. As an alternative, a neural network can be used to approximate possible nonlinear relationship between the nuisance trait and process data. The resulting neural network model for nuisance trait may be incorporated into the multidimensional 2PL model, so that the parameters for both components can be trained simultaneously.
Chapter 4: Computerized Adaptive Testing via Process Data

4.1 Introduction

Computerized Adaptive Testing (CAT) is an approach to adaptive test administration, which sequentially chooses the next items for an examinee based on their responses to previous items [76]. The goal of this approach is to efficiently and reliably measure proficiency, by selecting from a pre-calibrated item pool a subset of questions that can provide the most information on a particular examinee’s proficiency. Adaptive testing and multi-stage testing have been adopted for a wide variety of operational tests used for admissions (e.g., Graduate Record Examinations, Graduate Management Admissions Test), personnel selection and recruiting (e.g., Armed Services Vocational Aptitude Battery and IBM), professional licensure (e.g., National Council Licensure Examination for nursing, the Uniform CPA Examination), and large-scale assessments and accountability testing (e.g., National Assessment of Educational Progress, Smarter Balanced), see [44, 6, 77] for overviews of CAT.

CAT algorithms have been extensively studied in the measurement literature [see 78, 44, 79], and one typically consists of four critical components, namely a starting strategy for selecting initial questions to administer, a scoring rule for updating the examinee’s proficiency estimate throughout the test, an item selection algorithm for choosing subsequent question(s) based on the provisional proficiency estimate, and a termination strategy for determining when to stop administering new items. The item selection algorithm is the most crucial component that ensures the adaptivity of the test, as it determines which item will be given next based on the examinee’s responses observed so far. This leads to a personalized testing experience where each examinee receives a unique set of questions. Traditionally, scoring rules and item selection algorithms are based on final scores to test items, often established on an item response theory (IRT) or cognitive
diagnosis [see 80, 81] modeling framework: Items in the pool are precalibrated based on an item response model that relates the measured proficiency to the item final score (e.g., a binary score for correct or incorrect response) distribution. At each point in time during a test administration, the examinee’s proficiency estimate is updated based on the observed final responses to items administered so far. The adaptive selection algorithm then chooses the next question by optimizing an objective that, at least partially, depends on the expected decrease in measurement error on the measured proficiency from the chosen item’s final response. When the scoring rule is based on IRT, commonly adopted item selection strategies include the maximum Fisher information method [72], the global information approach based on Kullback-Leibler divergence [82], and $\alpha$-stratified methods [83, 84], and many others. Under cognitive diagnosis models, examples of item selection strategies include the use of Kullback-Leibler information [85, 86] and Shannon entropy [87, 86]. Recently, CAT item selection has also been approached from a recommender system perspective, with collaborative filtering and reinforcement learning applied to item selection [e.g., 88, 89, 90].

CAT algorithms based on final (binary, polytomous, or nominal) scores, which are commonly adopted for multiple-choice and short-answer tests, may see limited utility for complex tasks. Examples of complex tasks include constructed-response items and simulation-based assessments, which often demand more time and effort per item compared to multiple-choice or short-answer problems but, at the same time, collect richer proficiency evidence, such as the sequence of keystrokes to an open-ended question or the sequence of actions in pursuit of solving a simulation task. Application of final-score-based CAT to complex assessments is limited, for two reasons: To start, on-the-fly proficiency scoring of constructed responses is more challenging, given that the observed behavioral evidence on complex tasks is unstructured sequence data. In the event that constructed responses can be machine-scored in real-time to produce a final binary or polytomous score, the final score may not fully capture the rich proficiency information available from the observed sequence data. Especially in the initial phase of a CAT, when only a few items have been administered, the proficiency estimate based on final responses tends to be coarse, which further limits the adaptivity of the CAT to an examinee’s true proficiency: Two examinees with the same
final score but different observed test-taking processes (or open-ended responses) would receive the same proficiency estimate, leading to the selection of the same subsequent item, even if their true proficiency differs.

The current chapter presents a framework for incorporating test-taking process information into CAT with theoretical guarantees. Compared to a final score, such as a binary score that summarizes the degree of task accomplishment, process data can retain detailed information about individual differences, such as the examinee’s approach to reaching the final response and test-taking engagement. Extensive research has demonstrated the potential of process data for uncovering additional information: For instance, process data has been employed to develop measurement models for examinee latent traits or latent classes [91, 92, 93, 15] and to enhance proficiency scoring Chapter 2. Additionally, it has been used in cluster analyses to identify behavioral prototypes or problem-solving stages [94, 95, 96, 21, 97, 98], as well as to recognize behavioral attributes that can predict final performance [14, 99, 100, 101, 19]. The collection of test-taking process data is increasingly common for both simulation tasks and other computer-based exams [e.g., 102, 22].

The emergence of process data provides two main opportunities for improvements in CAT. Firstly, incorporating information from process data into the scoring rule has the potential to enhance the reliability of proficiency measurement. With both simulations and the PIAAC PSTRE assessment data, the process-incorporated scoring rule in Chapter 2 was found to achieve satisfactory accuracy even when only a small number of items are administered, where traditional measurement based on final responses tends to fall short. This improvement in proficiency measurement precision for short tests is particularly beneficial to CAT, which greatly relies on accurate initial proficiency estimates to select suitable items. As CAT begins utilizing the provisional proficiency estimate early on in the test for subsequent item selection, high measurement error in initial ability estimation has been consistently shown to carry over to the final proficiency estimate [e.g., 103, 104].

Secondly, the use of process data presents an opportunity to enhance item selection algorithms in CAT. Test-taking process data contains information not only about the measured proficiency but
also about other nuisance traits, such as the examinee’s interests or familiarity with the item format.

It is important to note that the expected amount of information an item can provide on the measured proficiency likely depends on the examinee’s true proficiency level as well as other nuisance traits. Assessing general proficiency is often inevitably realized by framing the item within a context that involves a specific content domain, theme, wording, and item format [38]. Even for equally capable examinees, differences in reception on these item-specific contexts can lead to variations in the amount of information an item can provide on the measured trait. For instance, lack of interest, lower perceived relevance, or technical barriers in working with specific item formats can result in lower test-taking engagement on an item, which reduces the amount of information an item can provide on the examinee’s true capabilities. In the presence of such interaction effects between nuisance factors and the measured proficiency on item informativeness, process data may provide information on the nuisance traits Chapter 3 that can inform the choice of subsequent items that optimize proficiency measurement accuracy.

The rest of this chapter is organized as follows. Section 4.2 presents a process-incorporated CAT framework, where real-time proficiency estimation, adaptive item selection, and model calibration are discussed. The performance of the proposed framework is evaluated through simulations in Section 4.3. Finally, in Section 4.4, we summarize the findings and provide some concluding remarks.

4.2 Computerized Adaptive Testing Algorithm via Proficiency Responses

In CAT, subsequent items are adaptively selected, based on the examinee’s observed data on items that have already been administered. Specifically, at a given time point $t$, the examinee has responded to a set of items $\mathcal{J}_1 = \{j_1, \ldots, j_t\}$. The remaining items in the pool, denoted by $\mathcal{J}_2 = \{1, \ldots, J\} \setminus \mathcal{J}_1$, make up the candidate item set. The item selection algorithm then chooses the next item for the examinee from $\mathcal{J}_2$ according to the examinee’s observed responses and, if available, process data on $\mathcal{J}_1$.

Under the IRT framework, item characteristics and final responses are assumed to follow a
specific IRT model, as elaborated in Section 1.1. Besides, item scores are assumed to be locally independent. As a consequence, the log likelihood of observed final responses $Y_{J_1}$ is additive, as represented by

$$l(\theta|Y_{J_1}) = \sum_{j \in J_1} Y_j \log P_j(\theta) + (1 - Y_j) \log(1 - P_j(\theta))$$ (4.1)

if the final responses $Y_j$ are binary. Note that in the CAT procedure, item parameters are assumed to be well-calibrated.

Over the years, IRT-based CATs have been extensively researched. Broadly speaking, these CAT algorithms fall into two categories: frequentist framework and the Bayesian framework. In the former, a prevalent approach involves the use of fisher information. If the MLE is adopted as the proficiency estimator, it is asymptotically normally distributed with variance inversely proportional to the sum of item Fisher information. Given the additive nature of fisher information, a consequence of the log likelihood’s additivity, the contribution of item $j$ to the information of MLE of $\theta$ is given by

$$I_j(\theta) = \frac{P_j'(\theta)}{P_j(\theta)[1 - P_j(\theta)]}$$ (4.2)

where $P_j'(\theta) = \frac{\partial P_j(\theta)}{\partial \theta}$. Thus one common strategy for minimizing the MSE is to select items that maximize the item Fisher information evaluated at the current proficiency estimate. Numerous adaptations of the objective functions have been proposed to overcome limitations and cater to different scenarios. The alternative category of CAT approaches is based on Bayesian methods. Essentially, the posterior distribution can be derived, and the precision of candidate items can be determined. A typical strategy is to minimize the expected variance, utilizing either a normal approximation or the true posterior distribution.

The collection of test-taking process provides another perspective to item selection in CAT. In many instances, the chance of solving a problem correctly, as well as how informative an item is for an examinee’s proficiency, depends on both the target proficiency and other nuisance traits. The same nuisance trait(s) could be involved across different items. For instance, for two items on the same topic, equally capable examinees with differing levels of interest in the topic may differ
in their attentiveness, which subsequently leads to differences the amount of information the items can provide on an examinee’s proficiency. At the same time, examinees’ attentiveness levels on the two items are likely correlated, as the two items involve the same topic. Process data on the first time, which can provide information on the nuisance (level of interest on the topic), may help determine how informative the second item is for a particular examinee.

The current work incorporates process information into trait estimation and item selection in CAT by using the process on each administered item \((X)\) to estimate an examinee’s target proficiency and nuisance trait(s), denoted as \((\hat{\theta}(X), \hat{\eta}(X))\). These item-level target and nuisance trait estimates from items administered so far are subsequently used to forecast proficiency estimator distributions on candidate items available for selection. Note that these proficiency estimators are continuous in most cases. IRT models that are commonly designed to handle categorical responses are not appropriate. Hence, a compatible probabilistic model is introduced to model the joint distribution of the proficiency and nuisance estimators across items.

In what follows, we introduce each key component to the process-incorporated CAT. Specifically, in Section 4.2.1, a heteroscedastic estimate model is used to illustrate how item-level proficiency and nuisance trait estimates are influenced by the underlying latent traits. Section 4.2.2 discusses how to further estimate the latent proficiency based on the proposed model, and function as real-time estimates. The CAT procedure, grounded on the heterogeneous response model and these real-time estimates, is presented in Section 4.2.4. Finally, Section 4.2.3 provides guidelines for item calibration, that is, in the presence of process data, how the process-incorporated item-level proficiency estimates and item-specific parameters in the proposed model (used for adaptive item selection) can be estimated.

4.2.1 Heteroscedastic Estimate Model

For each item \(j\), given the observed process features \(X_j\), consider the process-based proficiency estimator \(\hat{\theta}_j(X_j)\) and the process-based nuisance trait estimator \(\hat{\eta}_j(X_j)\). These item-level trait estimators, which are functions of the observed process \(X_j\), are assumed to be precisely calibrated
during the calibration stage, in the sense that their functional forms are known and designed to extract the information about the latent traits. An item $j$’s item-level process-based proficiency estimates are further assumed to depend on true proficiency $\theta$ and nuisance $\eta$ via a heteroscedastic estimate model:

$$
\begin{pmatrix}
\hat{\theta}_j \\
\hat{\eta}_j
\end{pmatrix}
\sim
N\left(
\begin{pmatrix}
\theta \\
\eta
\end{pmatrix},
\Sigma_j(\theta, \eta)
\right).
$$

(4.3)

Here $\Sigma_j(\theta, \eta)$ is the covariance matrix that may rely on the latent proficiency values of $\theta$ and $\eta$ to accommodate potential heterogeneity in the covariance matrix. Similar to the IRT models, we assume these estimators are locally independent given both $\theta$ and $\eta$. Understanding the existence of such heterogeneity involves recognizing that two examinees with differing traits might approach the same item in different ways as long as the target and nuisance proficiency influence the problem-solving process. Such variations could result in distinct distributions for the process-based item-level trait estimates. Note that if the covariance matrix does not depend on $(\theta, \eta)$, it reduces to the homoscedasticity case, where the precision of trait estimates on an item, in particular, the precision of the target proficiency estimate $(\hat{\theta}_j)$ is the same across examinees. In such a scenario, the item selection will become non-adaptive, as the variance is independent of the examinee’s ability or nuisance.

In the proposed heteroscedastic model, we make a further assumption that the covariance matrix follows a quadratic form. For simplification, let us denote $(\theta, \eta)^T$ as $\xi$. The covariance matrix is thus expressed as

$$
\Sigma_j(\theta, \eta) = A_j + \begin{pmatrix}
(\xi - b_{j1})^T C_{j11}(\xi - b_{j1}), (\xi - b_{j1})^T C_{j12}(\xi - b_{j2}) \\
(\xi - b_{j2})^T C_{j21}(\xi - b_{j1}), (\xi - b_{j2})^T C_{j22}(\xi - b_{j2})
\end{pmatrix}
$$

(4.4)

where $A_j, b_{j1}, b_{j2}$ and $C_{j11}, C_{j12} = C^T_{j21}, C_{j22}$ represent item parameters. Here, $A_j$ is a $2 \times 2$ covariance matrix intercept, $b_{j1}, b_{j2}$ are $2 \times 1$ column vectors acting as item proficiency intercepts and $C_{j11}, C_{j12}, C_{j22}$ are $2 \times 2$ matrix to scale different attributes of latent proficiency and manage
the interaction between $\theta$ and $\eta$. Note that the covariance matrix can also be expressed as

$$
\begin{pmatrix}
(\xi - b_{j1})^T, & 0^T \\
0^T, & (\xi - b_{j2})^T
\end{pmatrix}
\begin{pmatrix}
C_{j11}, C_{j12} \\
C_{j21}, C_{j22}
\end{pmatrix}
\begin{pmatrix}
(\xi - b_{j1}), & 0 \\
0, & (\xi - b_{j2})
\end{pmatrix}
$$

(4.5)

so as long as the matrix $C_j = \begin{pmatrix} C_{j11}, & C_{j12} \\
             C_{j21}, & C_{j22} \end{pmatrix}$ is positive definite, the overall covariance matrix is valid.

To comprehend the quadratic assumption, consider a simplified scenario focusing solely on $\theta$. In this case, the variance of the proficiency estimate $\hat{\theta}$ takes the form of $a + c(\theta - b)^2$. Then consider the 2PL IRT model, i.e.,

$$
P(Y = 1|\theta) = \frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}
$$

(4.6)

where $b$ is the item difficulty parameter and $a$ is the discrimination parameter. If we approximate the MLE with a normal distribution, it should have a mean as $\theta$ and variance as the reciprocal of the Fisher information. The latter can be derived as

$$
\frac{1}{I(\theta)} = \frac{1}{a^2} \left( e^{a(\theta - b)} + e^{-a(\theta - b)} + 2 \right).
$$

(4.7)

Employing a Taylor series expansion allows us to approximate the above as:

$$
\frac{1}{I(\theta)} = \frac{4}{a^2} + (\theta - b)^2 + O((\theta - b)^4)
$$

(4.8)

Therefore, the quadratic assumption can be interpreted as an approximation applicable in the case of continuous estimates, which provides substantial support for the practical use of the quadratic covariance matrix assumption.

It is also worth discussing the scenario where $C_{j12} = 0$, implying that the covariance between the proficiency estimate $\hat{\theta}_j$ and nuisance trait estimate $\hat{\eta}_j$ is a constant. Moreover, if the off-diagonal elements of $A_j$ are 0, $\hat{\theta}_j$ and $\hat{\eta}_j$ will be independent of each other. One could consider
\( \eta \) as representative of a preferred problem-solving strategy. In such a case, the observed target ability estimate may be independent of the problem-solving strategy. Nevertheless, the variance of \( \hat{\theta}_j \) could still be influenced by the problem-solving strategy, as captured by the \( C_{j11} \) matrix. Therefore, even if these estimates are independent, incorporating the nuisance trait can assist in estimating the target trait.

Furthermore, note that it is straightforward to generalize the heteroscedastic estimate model to accommodate multi-dimensional nuisance traits \( \eta \). To be specific, the \( m, n \) element of the covariance matrix \( \Sigma_j \) is given by

\[
\sigma_{jmn} = (\begin{pmatrix} \theta \\ \eta \end{pmatrix} - b_{jm})^T C_{jmn} (\begin{pmatrix} \theta \\ \eta \end{pmatrix} - b_{jn}).
\] (4.9)

### 4.2.2 Real-time Proficiency Estimation

The calibrated proficiency estimates \( \hat{\theta}_j \) and \( \hat{\eta}_j \) are valid estimators for corresponding latent trait as they are unbiased. However, they are not asymptotically efficient in general. From an intuitive perspective, these estimates can be viewed as sufficient statistics which summarize the attribute-related information derived from the process response. Under the framework of the heteroscedastic estimate model, MLE for \((\theta, \eta)\) is then adopted as the real-time latent traits estimators. To be specific,

\[
\hat{\theta}_t, \hat{\eta}_t = \arg \max_{j \in J_1} l_j(\theta, \eta | \hat{\theta}_j, \hat{\eta}_j).
\] (4.10)

The log likelihood function \( l_j(\xi | \hat{\xi}) \) has the form of

\[
-\frac{1}{2} \log \det(\Sigma_j(\xi)) - \frac{1}{2}(\hat{\xi}_j - \xi)^T \Sigma_j(\xi)^{-1}(\hat{\xi}_j - \xi).
\] (4.11)

Given that the covariance matrix also encompasses \( \xi \), there typically isn’t an analytical formula for the MLE. However, considering that the determinant of the covariance matrix is a polynomial of \( \xi \) and the precision matrix is also a polynomial fraction, the derivative of the log likelihood
function can be represented as a polynomial fraction of $\xi$. This allows the MLE to be numerically and efficiently found.

Last but not the least, Bayesian estimators could serve as viable alternatives to the MLE, assuming a suitable prior for $(\theta, \eta)$ has been established. By incorporating additional information from the prior, Bayesian estimators, such as the EAP and MAP commonly used in IRT models, have the potential to enhance proficiency estimates. However, in this dissertation, we primarily concentrate on employing the MLE as the estimator, leaving the exploration of alternative approaches to future research endeavors.

4.2.3 Model Calibration

Similar to item calibration before operational use in the IRT-based CAT setting, there are two critical components that require calibration in the proposed process-incorporated CAT algorithm, (1) the process-incorporated proficiency estimate and nuisance trait estimate, and (2) the item parameters in heteroscedastic estimate model employed for adaptive item selection. This section introduces a feasible item calibration procedure that calibrates one item at a time.

Suppose we are presented with a new item that needs calibration in accordance with the established items in the item bank. Then for each examinee $i = 1, \ldots, I$, the process responses $S_i$ are recorded and the underlying traits are estimated by existing items as $(\theta_i, \eta_i)$. The calibration procedure is outlined as follows:

**Procedure 2.**

1. Extract the process features $X_i$ from the observed sequence $S_i$ by methods discussed in Section 1.3.

2. Regress the latent traits $\theta_i$ and $\eta_i$ on the process features $X_i$ to obtain the proficiency estimate $\hat{\theta}(X)$ and nuisance trait estimate $\hat{\eta}(X)$.

3. Given the observed latent traits estimate $\hat{\theta}(X_i), \hat{\eta}(X_i)$ and the mean of heteroscedastic estimate model $(\theta_i, \eta_i)$, obtain MLE for the item parameters $A$, $b$s and $C$s.
4.2.4 Adaptive Item Selection

Similar to the paradigm in the IRT framework, MLE is adopted as the proficiency estimator and item responses are assumed to be locally independent. Therefore, the item is selected to maximize the Fisher information among the candidate item set \( J_2 \). For a normal distribution, it can be easily verified that the Fisher information matrix equals the precision matrix, so the Fisher information about the target trait \( \theta \) is found as the top left element of this matrix, which has the form of

\[
\frac{1}{\sigma_{11}(\theta, \eta)} - \frac{\sigma_{12}^2(\theta, \eta)}{\sigma_{22}(\theta, \eta)}
\]

(4.12)

where \( \Sigma(\theta, \eta) = \begin{pmatrix} \sigma_{11}(\theta, \eta), \sigma_{12}(\theta, \eta) \\ \sigma_{21}(\theta, \eta), \sigma_{22}(\theta, \eta) \end{pmatrix} \). Note that the asymptotic variance of MLE is the reciprocal of the Fisher information, and the additive part of the Fisher information coincides with the reciprocal of conditional variance of \( \theta \) given \( \eta \). Hence, as long as \( \hat{\theta} \) depends on \( \hat{\eta} \), the inclusion of the nuisance trait will enhance the precision of the target trait estimation.

Assuming that the item parameters in the heteroscedastic estimate model are calibrated as discussed in Section 4.2.3, we outline the process-based CAT algorithm as follows:

**Procedure 3.**

*For each examinee,*

1. Select the first item randomly. The selected item is denoted as \( j_1 \).

2. For each time point \( t = 1, 2, 3, \ldots \),
   1. Extract the process features \( X_{jt} \) from the observed sequence \( S_{jt} \) on item \( j_t \).
   2. Calculate the latent trait estimates on item \( j_t \), \( \hat{\theta}_{jt} \) and \( \hat{\eta}_{jt} \), using calibrated estimate functions.
   3. Obtain the MLE as \( \hat{\theta}_t, \hat{\eta}_t \) defined in Equation (4.10).
3. For each candidate item \( j \in J_2 = \{1, \ldots, J\} \setminus J_1 \), compute the Fisher information Equation (4.12) by plugging in the current estimate \( \hat{\theta}_{jt} \) and \( \hat{\eta}_{jt} \).
5. Select the item \( j \) that maximizes the Fisher information, denoted as \( j_{t+1} \), and add the item to \( J_t \).

6. Repeat the process until the termination rule has been reached, e.g., after administering a desired number of items.

### 4.3 Simulation Studies

#### 4.3.1 Experiment Settings

In this simulation study, we consider a two-dimensional latent trait \((\theta, \eta)\), where \( \theta \) denotes the primary problem-solving ability, and \( \eta \) represents the nuisance ability. To generate the data, we drew independent random variables from a standard normal distribution for both \( \theta_1, \ldots, \theta_I \) and \( \eta_1, \ldots, \eta_I \) of \( I = 1000 \) respondents. The dataset used in this study comprises 200 items, a number ten times the intended test length of 20.

The impact of covariance structure on the performance of the CAT algorithm is assessed across three distinct scenarios, each differing in the method of item parameter simulation. In the first scenario, no constraints are placed on the item parameters. For each item \( j \), the item parameter \( A_j \) is generated from a Wishart distribution using the identity matrix as the scale matrix parameter. The elements of the intercepts \( b_s \) are sampled from a uniform distribution. The discrimination matrix parameters \( C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \) are likewise generated from a Wishart distribution with a dimension of 4 to ensure the covariance matrix \( \Sigma_j(\theta, \eta) \) is positive definite.

The second scenario contemplates a special case where \( \hat{\theta}_j \) is independent of \( \hat{\eta}_j \), i.e., \( C_{12} = 0 \) for any item \( j \) and off-diagonal elements of \( A = 0 \). The generation of item parameters proceeds similarly with the first scenario. Diagonal elements of \( A \) are simulated from the chi-square distribution. For parameters \( C_s \), both \( C_{11} \) and \( C_{22} \) are generated from a Wishart distribution with dimensions set to 2.

In the third scenario, we consider another special case where the marginal distribution of \( \hat{\theta} \) is irrelevant to the nuisance trait \( \eta \), i.e., \( \text{Var}(\hat{\theta}) = a + c(\theta - b)^2 \). This scenario investigates performance
outcomes when only $\theta$ is considered, thereby omitting some dependent nuisance traits. To ensure the desired variance structure and uphold the positive definiteness, we initially generate a $3 \times 3$ positive definite matrix from the Wishart distribution. This is then expanded to a $4 \times 4$ matrix with the second row and column set to zero, which is designated as parameter $C$. It is straightforward to verify that, in this case, the variance of $\hat{\theta}$ is unrelated to $\eta$.

Upon establishing the item parameters, latent trait estimates are simulated from the normal distribution, using the mean and covariance matrix outlined in the heteroscedastic estimate model. To ensure the validity of the results, each scenario is replicated ten times.

4.3.2 Comparison and Evaluations

In the present study, we conduct a comparative analysis of four different item selection algorithms:

1. Random Selection: In this approach, items are chosen at random from the pool of 200 items for each examinee.

2. Real-Time Process-Based Selection: Items are selected according to the procedure outlined in Procedure 3, which relies on real-time latent trait estimates.

3. True Proficiency Process-Based Selection: This method is identical to the real-time process-based selection, but it substitutes real-time latent trait estimates with true values of $\theta$ and $\eta$ when computing the Fisher information. This method is not feasible in practice; however, its evaluation provides an upper performance limit.

4. Process-Based Selection Focused on $\theta$: This approach also follows the procedure delineated in Procedure 3, but it only incorporates $\theta$ in the heterogeneous model. This algorithm is exclusively utilized in the third scenario, where the marginal distribution of $\hat{\theta}$ is solely dependent on $\theta$.

For all four algorithms, the first selected item is randomly selected. These item selection procedures are then repeated until the desired test length of 20 is achieved.
To compare true proficiency recovery, two criteria were considered including root mean square error (RMSE) and Kendall’s $\tau$ between the estimated and the true $\theta$. At each test length $t = 2, \ldots, 15$, RMSE was calculated as follows:

$$RMSE(\hat{\theta}^t) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{\theta}_i^t - \theta)^2}.$$  \hfill (4.13)

These criteria were also applied to evaluate the performance of the four item selection algorithms described in Section 4.3.2.

To evaluate item exposure and pool utilization, the mean of test overlap rates ($tor_{ii'}$) [105] between all possible pairs of examinees $i$ and $i'$ is adopted,

$$\bar{tor} = \left(\frac{I}{2}\right)^{-1} \sum_{i=1}^{I-1} \sum_{i'=i+1}^{I} tor_{ii'}$$  \hfill (4.14)

where $tor_{ii'}$ is computed as the number of common items between a pair of examinees divided by test length.

4.3.3 Results

Figure 4.1 presents the results on the recovery of true proficiency in the first scenario based on the three methods, specifically, random selection, maximal Fisher information selection with real-time estimates and maximal Fisher information selection with true latent traits. The simulation outcomes suggest that the proposed CAT algorithm significantly surpasses the random selection in terms of both RMSE and Kendall’s $\tau$. A closer examination of the performance revealed that the CAT algorithm with real-time estimates shows comparable efficacy to random selection with twenty items, even when only eight items are administered, as per both metrics. When compared to the theoretical upper limit provided by the CAT algorithm with true latent traits, the gap progressively diminishes as more items are responded to. This observation aligns with our intuition, considering that the initial phase estimates may not be particularly accurate, potentially leading
Figure 4.1: Comparison of CAT algorithms in Scenario 1
to suboptimal item selection. However, as more items are administered, the real-time estimates increasingly approach the true values, subsequently reducing item selection errors.

In the second scenario where $\hat{\theta}$ and $\hat{\eta}$ are independent, Figure 4.2 clearly demonstrates that the proposed CAT algorithm maintains superior performance compared to the random selection method. Although the incorporation of the nuisance attribute does not reduce the asymptotic variance of the MLE, it enables a more precise variance structure and thus leads to a more accurate variance estimation. Nevertheless, it’s noticeable that the gap between the CAT using real-time estimates and the theoretical upper bound is larger than in the first scenario. This is intuitively comprehensible since the estimates exhibit a greater variance and are less accurate. Consequently, the deviation from the true value is more substantial compared to the first scenario, requiring more time to narrow the discrepancy.

The performance comparison for the item selection algorithms in the third scenario is illustrated in Figure 4.3, introducing an additional item selection algorithm that solely considers $\theta$. Given that the marginal distribution depends exclusively on $\theta$, there’s no model misspecification in this case. Its performance aligns closely with that of the random selection using the full model in terms of RMSE and displays superior performance in terms of Kendall’s $\tau$. This suggests that the CAT algorithm remains effective even when some latent traits are omitted. Besides, the performance variation is larger due to the absence of consideration for $\eta$ resulting from a larger variance in the estimate. Moreover, since $\theta$ and $\eta$ are dependent in this scenario, the full model surpasses the algorithm that includes only $\theta$, as $\eta$ can provide supplemental information.

Lastly, for a comparison of item pool usage for different item selection algorithm, Figure 4.4 presents the average test overlap rates across the 10 replications. Note that the random selection method provides a lower bound of 0.1. The proposed CAT algorithm, when incorporating the nuisance attribute $\eta$, delivered better outcomes in terms of item exposure, as the average TOR was significantly lower than that without considering $\eta$. This aligns with the intuition that the variation in respondents should increase with more latent attribute dimensions taken into consideration, leading to a more diversified item selection. When only $\theta$ is taken into account, respondents are more
Figure 4.2: Comparison of CAT algorithms in Scenario 2
Figure 4.3: Comparison of CAT algorithms in Scenario 3
likely to encounter identical items—even in instances where their nuisance traits $\eta$ vary—leading to an elevated average TOR.

Figure 4.5 displays the distribution of item exposure rate for a single replication, indicating that the proposed method with full model had a more balanced exposure rate than that with only target trait, and by comparing the number of items ever selected by different methods, the proposed method selected approximately twice as many items as that without consideration of nuisance trait. Therefore, the proposed CAT algorithm alleviated the issues of over-exposure and under-exposure without explicit design. These findings demonstrate that the proposed algorithm is not only more efficient and accurate in selecting items, but also resulted in better item pool usage.

4.4 Conclusion

This chapter introduces a novel method that integrates process information into proficiency estimation and CAT item selection. A heteroscedastic estimate model is proposed to model the profi-
Figure 4.5: Comparison of exposure rate

ciency estimate and the nuisance trait estimate. A corresponding CAT algorithm is also presented. Simulation results indicate that the maximal Fisher information algorithm for item selection, when used in conjunction with the the heteroscedastic estimate model, outperforms the other algorithms considered in terms of various evaluation metrics of true proficiency recovery and item exposure control.

The process-incorporated CAT algorithm has several practical implications, particularly when it comes to assessments that consist of complex tasks such as open-ended problems or simulation tasks. Traditionally, CAT has seen limited applications in these areas due to the challenges associated with real-time scoring, the time-consuming nature of each task, and the high cost of item pool development for complex tasks. The proposed process-incorporated CAT algorithm partially addresses these concerns. First, by introducing a heteroscedastic estimate model based on observed sequence data, it allows for real-time proficiency scoring based on the test-taking process and, similarly, constructed responses to open-ended problems. Second, by incorporating the
additional information on both the measured proficiency and nuisance traits in the process data, the process-incorporated scoring and CAT algorithms substantially reduce the test length required to achieve a given level of measurement precision. Third, with item design, review, and field testing as one of the leading costs in test development, the proposed method achieved more balanced item pool usage. Consequently, this new approach has the potential to expand the applicability of CAT in complex domains, ultimately leading to more accurate, efficient, and ecologically valid assessments of individuals’ abilities and skills.

As an initial study exploring the integration of process-incorporated estimation and computerized adaptive testing, the present study has several limitations. Firstly, heteroscedastic estimate model relies on the assumption of the covariance matrix structure, which may not always hold. This may lead to overestimation or underestimation of the Fisher information and result in incorrect item rankings. Secondly, the response processes were transformed into process features using generic methods, rather than methods specifically designed for computerized adaptive testing. This means that important information may have been lost during the transformation process, which could negatively impact the performance of the adaptive testing system.

To address these limitations, several avenues for future research could be explored. First, more complex covariance structure could be investigated as an alternative to the current heteroscedastic estimate model to better capture the complex relationship between the Fisher information and latent traits. Secondly, it may be possible to develop an end-to-end solution that directly accepts response processes as inputs and predicts the objective. This would eliminate the need for intermediate feature extraction and prevent information loss, at the same time allowing for the use of inputs with variable lengths. Finally, the ideas from recommendation systems could be leveraged to adopt alternative objectives, such as mean reciprocal rank, mean average precision, and normalized discounted cumulative gain, which place more emphasis on rankings. These potential research directions could further advance the integration of process-incorporated estimation and computerized adaptive testing.
References


96


Appendix A: Proofs of Theorem 1 and Theorem 2

To prove Theorem 1, we establish the following lemma.

Lemma 1. Let $X$ be a nonconstant random variable, and $f(\cdot)$ and $g(\cdot)$ be strictly increasing functions. Suppose that $f(X)$ and $g(X)$ have finite second moments. Then $\text{Cov}(f(X), g(X)) > 0$.

Proof of lemma 1. Let $Y$ be an independent and identically distributed (i.i.d.) copy of $X$. It is easy to verify the following identity

$$
\text{Cov}(f(X), g(X)) = \frac{1}{2} E [(f(X) - f(Y))(g(X) - g(Y))]. \tag{A.1}
$$

Clearly, for any $x$ and $y$, $(f(x) - f(y))(g(x) - g(y)) \geq 0$, and "='' holds if and only if $x = y$. Since $P(X \neq Y) > 0$, the right-hand side of equation (A.1) must be positive. \hfill \Box

Proof of Theorem 1. By Assumption A2 (local independence),

$$
T_{X_{-j}} = E \left[ \hat{\theta}_{y_j}|X_{-j} \right] = E \left[ E \left[ \hat{\theta}_{y_j}|X_{-j}, \theta \right] \big| X_{-j} \right] = E \left[ E \left[ \hat{\theta}_{y_j}|\theta \right] \big| X_{-j} \right] = E \left[ m_j(\theta)|X_{-j} \right].
$$

Due to Assumption A3 (exponential family), the posterior distribution of $\theta$ given $X_{-j}$ depends on $X_{-j}$ only through the sufficient statistic $T_j(X_{-j})$. In fact,

$$
T_{X_{-j}} = E \left[ m_j(\theta)|X_{-j} \right] = G_j(T_j(X_{-j})),
$$

where $G_j(t) = E \left[ m_j(\theta)|T_j(X_{-j}) = t \right]$. Furthermore, by making use of the exponential family form in Assumption A3 and the simple exchange of order of differentiation and integration, we
can show that
\[ G'_j(t) = \text{Cov} \left[ m_j(\theta), \eta_j(\theta) \mid T_j(X_{-j}) = t \right]. \]

Since both \( m_j \) and \( \eta_j \) are strictly monotone, Lemma 1 implies that \( G'_j(t) \) is strictly positive or negative for all \( t \) and, therefore, \( G_j \) is strictly monotone. In other words, there is a one-to-one mapping between \( T_{X_j} \) and \( T_j(X_{-j}) \).

\[ \square \]

**Proof of Theorem 2.** From Theorem 1, we know that \( T_{X_{-j}} \) is a sufficient statistic of \( X_{-j} \) for each \( j \). Since \( \hat{\theta}_Y \) is a function of \( Y \) and \( \sigma(Y_{-j}) \subseteq \sigma(X_{-j}) \), the conditional distribution \( \hat{\theta}_Y \mid T_{X_{-j}}, Y_j \) is free of \( \theta \). Therefore, we have \( E[\hat{\theta}_Y \mid T_{X_{-j}}, Y_j, \theta] = E[\hat{\theta}_Y \mid T_{X_{-j}}, Y_j] = \hat{\theta}_{X_{-j}} \). It follows from the well-known Rao-Blackwell Theorem [41] that \( \hat{\theta}_{X_{-j}} \) reduces the conditional variance and

\[ E[(\hat{\theta}_{X_{-j}} - \theta)^2 \mid \theta] \leq E[(\hat{\theta}_Y - \theta)^2 \mid \theta] \]

holds for every \( j \) and \( \theta \). By Cauchy-Schwarz inequality, we get

\[ E[(\hat{\theta}_X - \theta)^2 \mid \theta] \leq E[\frac{1}{J} \sum_{j=1}^{J} (\hat{\theta}_{X_{-j}} - \theta)^2 \mid \theta] \leq E[(\hat{\theta}_Y - \theta)^2 \mid \theta]. \]

\[ \square \]
Appendix B: Proof of Theorem 3

Proof of Theorem 3. We simplify the notations as follows. Assume for focal group

\[ X_i \overset{i.i.d.}{\sim} f_f(X_i), \quad i = 1, \ldots, N_f \]  \hspace{1cm} (B.1)

while for reference group

\[ X_i \overset{i.i.d.}{\sim} f_r(X_i), \quad i = N_f + 1, \ldots, N_f + N_r. \]  \hspace{1cm} (B.2)

Under null hypothesis, suppose

\[ P(Y_i = 1|X_i, \beta_0) \triangleq p_0(X_i) = \frac{e^{\beta_0^T X_i}}{1 + e^{\beta_0^T X_i}} \]  \hspace{1cm} (B.3)

for both groups. Let \( \hat{\beta}_f \) and \( \hat{\beta}_r \) denote the estimation respectively and

\[ P(Y_i = 1|X_i, \hat{\beta}_f) \triangleq p_f(X_i) = \frac{e^{\hat{\beta}_f^T X_i}}{1 + e^{\hat{\beta}_f^T X_i}} \]  \hspace{1cm} (B.4)

and

\[ P(Y_i = 1|X_i, \hat{\beta}_r) \triangleq p_r(X_i) = \frac{e^{\beta_r^T X_i}}{1 + e^{\beta_r^T X_i}} \]  \hspace{1cm} (B.5)

Then the likelihood function for the focal group has the form

\[ l(\beta) = \sum_{i=1}^{N_f} \log f_f(X_i) = \sum_{i=1}^{N_f} Y_i \log p_f(X_i) + (1 - Y_i) \log(1 - p_f(X_i)) \]  \hspace{1cm} (B.6)
and we have the Fisher information matrix has the form of

\[- \mathbb{E} \frac{\partial^2 l}{\partial \beta^2} = \sum_{i=1}^{N_f} \frac{e^{\beta^T X_i}}{(1 + e^{\beta^T X_i})^2} X_i X_i^T. \tag{B.7}\]

Therefore, by Central Limit Theorem and Law of Large Number,

\[\sqrt{N_f}(\hat{\beta}_f - \beta_0) \xrightarrow{d} N(0, I_f(\beta_0)^{-1}) \tag{B.8}\]

where

\[I_f(\beta_0) = \mathbb{E}_{X \sim f_f(X)} \frac{e^{\beta^T X}}{(1 + e^{\beta^T X})^2} X X^T \tag{B.9}\]

and

\[\frac{1}{N_f} \sum_{i=1}^{N_f} \frac{e^{\beta^T X_i}}{(1 + e^{\beta^T X_i})^2} X_i X_i^T \xrightarrow{a.s.} I_f(\beta). \tag{B.10}\]

Similarly

\[\sqrt{N_r}(\hat{\beta}_r - \beta_0) \xrightarrow{d} N(0, I_r(\beta_0)^{-1}) \tag{B.11}\]

where

\[I_r(\beta_0) = \mathbb{E}_{X \sim f_r(X)} \frac{e^{\beta^T X}}{(1 + e^{\beta^T X})^2} X X^T \tag{B.12}\]

and note that they are independent.

Next we come back to the definition of empirical $L^2$ distance $\hat{d}_j$. Based on Taylor’s expansion,
the difference between estimated probability

\[ p_f(X) - p_r(X) = \frac{e^{\beta_f^T X}}{(1 + e^{\beta_f^T X})^2}X^T (\hat{\beta}_f - \beta_0) - \frac{e^{\beta_r^T X}}{(1 + e^{\beta_r^T X})^2}X^T (\hat{\beta}_r - \beta_0) \]

\[ + o(\hat{\beta}_f - \beta_0) + o(\hat{\beta}_r - \beta_0) \tag{B.13} \]

\[ = \frac{e^{\beta_f^T X}}{(1 + e^{\beta_f^T X})^2}X^T (\hat{\beta}_f - \beta_0) - \frac{e^{\beta_r^T X}}{(1 + e^{\beta_r^T X})^2}X^T (\hat{\beta}_r - \beta_0) \]

\[ + o(\hat{\beta}_f - \beta_0) + o(\hat{\beta}_r - \beta_0) \tag{B.14} \]

\[ = \frac{e^{\beta_f^T X}}{(1 + e^{\beta_f^T X})^2}X^T (\hat{\beta}_f - \hat{\beta}_r) + o_p(1) \tag{B.15} \]

Let \( Z = (p_f(X_1) - p_r(X_1), \ldots, p_f(X_N) - p_r(X_N))^T \) to denote the difference on all examinees.

Then

\[ N \hat{d}_j^2 = Z^T Z = (\hat{\beta}_f - \hat{\beta}_r)^T \left( \sum_i \frac{e^{\beta_f^T X_i}}{(1 + e^{\beta_f^T X_i})^2} X_i^T X_i \right) (\hat{\beta}_f - \hat{\beta}_r) \tag{B.16} \]

\[ = \sqrt{N}(\hat{\beta}_f - \hat{\beta}_r)^T \left( \frac{1}{N} \sum_i \frac{e^{\beta_f^T X_i}}{(1 + e^{\beta_f^T X_i})^2} X_i^T X_i \right) \sqrt{N}(\hat{\beta}_f - \hat{\beta}_r) \tag{B.17} \]

where

\[ \frac{1}{N} \sum_i \frac{e^{\beta_f^T X_i}}{(1 + e^{\beta_f^T X_i})^2} X_i^T X_i \xrightarrow{a.s.} \mathbb{E} \left[ \frac{e^{\beta_f^T X}}{(1 + e^{\beta_f^T X})^2} X^T X \right] \]  \tag{B.18} \]

Hence as long as \( \frac{N_f}{N_r} \) converges to a non-zero ratio, \( \sqrt{N}(\hat{\beta}_f - \hat{\beta}_r) \) will converge to a normal random variable and \( N \hat{d}_j^2 \) will converge to a generalized chi-square distribution.